INFORMATION PERSPECTIVE ON TURBULENCE

by

Rory Thomas Cerbus

B.S. Honors Physics, Purdue University, 2008

M.S. Physics and Astronomy, University of Pittsburgh, 2011

Submitted to the Graduate Faculty of
the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Rory Thomas Cerbus

It was defended on

November 4, 2014

and approved by

Walter Goldburg, Department of Physics and Astronomy, University of Pittsburgh

Pinaki Chakraborty, Fluid Mechanics Unit, Okinawa Institute of Science and Technology

Xiao-Lun Wu, Department of Physics and Astronomy, University of Pittsburgh

David Jasnow, Department of Physics and Astronomy, University of Pittsburgh

Vittorio Paolone, Department of Physics and Astronomy, University of Pittsburgh

Stephen Garoff, Department of Physics, Carnegie Mellon University

Dissertation Advisors: Walter Goldburg, Department of Physics and Astronomy, University of Pittsburgh,

Pinaki Chakraborty, Fluid Mechanics Unit, Okinawa Institute of Science and Technology

INFORMATION PERSPECTIVE ON TURBULENCE

Rory Thomas Cerbus, PhD

University of Pittsburgh, 2014

Turbulence is a complex system. Information theory [1, 2, 3, 4], and the more recently developed computational mechanics [5], provide quantitative tools for quantifying this "complexity" in terms of predictability. This thesis is concerned with using these tools with experimental turbulence data. The focus will be on two-dimensional turbulence [32, 33], but other (mostly turbulent) systems are treated as well. The role of information theory in physical systems in general will also be discussed. This framework is general and easily applicable to almost any other kind of system, as will be shown. A special transition is observed in which structureless fluctuations become cascade turbulence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I am extremely grateful to the many people who have taken the time to teach me good physics. First and foremost is, of course, Walter Goldburg. Walter took me in, as he does with so many, and patiently led me along. We spent many hours discussing what experiments to do, thinking about how to do them and finally arguing about what on earth our results could possibly mean. I didn't know how fun physics could be until he taught that to me. Some people look forward to the weekend as an escape from work. Walter would often call and email me on the weekend after thinking of a new idea or reading an interesting paper. He taught me to love thinking about physics (nearly) all the time.

Perhaps more than teaching me physics, I learned about life from him. I have never met a person more gracious and patient than Walter. I certaintly benefitted from his kindness, and I hope that I can share the same generosity of spirit to those I know and meet.

There have been quite a few other good scientists who have helped me along. Xiao-Lun Wu was the first physicist that I had the opportunity to work for. I'm still not sure why he hired me (twice), but I am grateful. Few researchers spend as much time as he and Walter do talking and arguing with their students. I learned a lot from him and fell in love with turbulence after working in Xiao-Lun's lab.

Stephen Garoff encouraged me at the beginning of my time as a graduate student. Steve's enthusiasm is infectious and working with him was a lot of fun. Solving a problem and publishing an article about it with him was a big boost to my suffering confidence. I only regret that I didn't have the opportunity to work with him more.

Dan Boyanovsky is the best teacher that I know. After taking his general relativity class, I entered into every class he offered that I was able to attend. His understanding of physics is thorough, and his willingness to explain things to me and others for hours after the lectures is inspiring. Those chats also left me feeling encouraged that I was beginning to understand the physical world around me in a deeper way. While many scientists tend to have tunnel vision and only find their own field of any interest, Dan loves and knows about an incredibly wide range of physics. I was able to get out of my tunnel a little bit through his help.

Pinaki Chakraborty has helped me in more ways than I can recall. I still don't know why he was willing to co-advise me, fund me and spend so much time fighting with bureaucracy for me, but I think it is because he is a generous, kind person. It was very helpful for me to learn about science, especially turbulence, from his perspective. He cuts through the clutter and tries to find the true (and hopefully simple) picture of how nature works. As with Walter, I have benefitted tremendously from discussions with Pinaki about physics

and about life.

I am grateful to my thesis committee for the time they spent reviewing my work and giving me advice. I am especially thankful for David Jasnow's thorough and critical reading of my thesis. I am also thankful for the other good teachers that I had while at Pitt, such as Bill Layton in the math department, Paul Shepard, Ralph Roskies, Bob Devaty and although he wasn't my teacher, Adam Leibovich.

I have benefitted greatly from the time spent with many of Walter's former students: Jason Larkin, Tuan Tran, Penger Tong, Rob Cressman and Mahesh Bandi. Mahesh has especially helped me out and taught me a lot of interesting physics. He very generously taught both Walter and I the basics of information theory and got us started on our quest to understand and use it.

Graduate school would have been much less enjoyable if it were not for the many other students I met and befriended here at Pitt. I am happy to have met Chitra, Mike, Alex, Kevin, Kevin, Elliot, Advait, Louis, Jen, Jun, Li, Jing and others. I am especially proud to have become friends with Sheng Huang, his wife Jane and son Christopher. Stefanus deserves a special mention. We never actually worked on the same project, despite working in the same lab for almost five years, but we talked about a lot of physics and life. I will never know half of what Stefanus does, nor understand it nearly half as well. He is a special, gifted person.

I will thank Chien-chia Liu here at the transition between thanking scientists and family. I didn't know how important Chien-chia, his wife Jui-yin and their daughter Mika would become in my and my family's life. Chien-chia has so much fun with life and physics. His energy is infectious and his patience inhuman. I am grateful for their friendship and encouragement these past few years.

The staff in the physics department here at Pitt have made all the steps and stages here smooth and painless. Leyla Hirschfeld has saved me from one disaster (always of my making) after another. She also first introduced Walter and I to each other. Lori Neu has been a good friend and a fun glass blowing teacher. It was good to have such a wonderful neighbor.

My parents and siblings have been a constant encouragement and support for as long as I can remember. I can never really be discouraged, because I know that my Mom loves me and will always be there for me. She is so strong. Dad first got me interested in physics and science. I always wanted to be like him, not just by getting a science degree but also by being so selfless and thoughtful. I received my PhD in five and a half years. I will spend the rest of my life learning to be a servant like he was.

My wife's family has helped me and my family out in an uncountable number of ways. They are constantly giving of what they have to help us, and I can never repay them. I have especially enjoyed the long chats with Dad Stahl about science and learning about teaching and chemistry from him.

My wife Bethany deserves her own degree for seeing us through this. She has given up many things and worked tirelessly to support me. I can never repay her, but I will try. I can't imagine having done this without her and I wouldn't have wanted to anyways.

Wendell encouraged me without knowing it. I can't help but smile and laugh when I see him and think

about him. He is a great joy in my life. His future siblings, such as Tako, will also give me greater joy than anything else.

I also want to thank my Lord, from whom all these blessings flow.

1.0 INTRODUCTION

The purpose of this thesis is to use information theory [4, 2], and its near kin computational mechanics [5], in the study of turbulence. Information theory is concerned with messages and their content, but it has found extensive use in physical systems. In the specific case here, we use it with the idea that our experimental data stream is turbulence's message to us. Computational mechanics goes further by determining how this information is shared and transferred in the message and by extension in the original system. Put another way, it treats the system like a computer and characterizes its computing in terms of information theory.

Turbulence has long been regarded as a difficult, interesting and unsolved problem [35]. It is unsolved in the sense that its main features have not been derived starting from a few fundamental principles. Because of this, and the richness of its behavior, many tools and ideas have been applied to it or born from it.

In this thesis we will focus almost exclusively on two-dimensional (2D) turbulence. Although we will touch on some fundamental issues regarding the "classical" theory of two-dimensional turbulence, our real concern will be with an information-theoretic treatment.

Two-dimensional turbulence is similar to three-dimensional (3D) turbulence but there are distinct differences. Both have theories that involve cascades as central players. A cascade is a statistical description[1] of the process of some quantity (*e.g.* energy) being transferred from one scale to another. Although the mechanism is not always clear, a common picture is that this occurs through the breaking up or combination of eddies, *i.e.*, turbulent motion at specific scales [35].

In 3D turbulence there is one cascade. As the Reynolds number Re, the parameter characterizing the turbulence's strength, becomes large, the viscous dissipation of energy becomes negligible at all but the smallest scales. Any energy put into the system will then travel through the break up of eddies until it reaches these small scales.

In 2D turbulence eddies tend to combine instead of break up, and due to the reduced dimensionality there is no vortex stretching [33, 32]. This is the process whereby vorticity can be strengthened by the elongation of the fluid element in the direction perpendicular to its plane of rotation. This is just what figure skaters do to increase their rotational speed by bringing their appendages closer to their bodies. The absence of this effect means that in addition to energy, enstrophy (squared vorticity[2]) is also conserved for the scales where dissipation is negligible. So in 2D turbulence, we have an inverse energy cascade to large scales and a direct

[1] Information theory is also statistical.

[2] Actually, any power of vorticity greater than 1 is conserved but 2 is easier to handle.

enstrophy cascade to small scales.

At this point in a manuscript one often finds a justification for studying 2D turbulence based on its applicability to the 2D nature of atmospheric turbulence, stratified oceanic turbulence, *etc.* We will not do so here. Rather, we follow in the spirit of G. H. Hardy, who argued for the study of pure mathematics because it is beautiful, interesting and deep [126].

We will study some of the general cascade ideas about 2D turbulence and also the deviations from their simple predictions. This is the study of intermittency. We find that although many have supposed this area of research to be closed (see the recent review article [33] and the references therein), our experiments suggest otherwise. A serious practical issue raised by this finding is the need to understand the differences between simulations and experiments of 2D turbulence.

The primary focus will be information theory. Here we will focus on using information theory and computational mechanics to characterize the degree of both unpredictability and predictability of our 2D turbulent flow. We will also take a short look at intermittency, the deviation from the basic theory of turbulence [51, 35].

Information theory's primary player is the entropy rate (density) h [4]. For now let us think of this (unpredictable) information as the number of "yes" or "no" questions we need to ask on average to guess any part of a data stream. Our h is the information we get from peeking at any data point having knowledge of all previous data points. We see how this is an excellent quantifier of unpredictability, as it will be large if we have no idea of what comes next and small otherwise. We will also use the excess entropy E. This is the average number of "yes" or "no" questions we no longer have to ask once we have looked at all previous data [21, 22].

For computational mechanics we have C. This is the average number of "yes" or "no" questions we need to ask to predict what will happen in the future (statistically) [5, 22]. Although this may seem similar to h, we will show that it is quite different.

Our main results focus on h and C (and the related E). In short, because of the strong correlations inherent in a cascade, the unpredictability decreases while the predictability increases as the turbulence strengthens. Because the cascade has this effect on the behavior of h, E and C, the absence of a cascade is easily detectable. Indeed, we are able to observe a new kind of transition where "noisy" velocity fluctuations become more structured with the development of a cascade.

2.0 INFORMATION THEORY

We will begin by motivating and defining the information tools used in this study of turbulence. The concept of information used here is slightly different from the everyday definition. In its essentials, however, it is the same. For us, information is "...what you don't already know." [1]. The statement "it will snow in Okinawa" contains a lot of information, while being told "it will be cloudy in Pittsburgh" holds very little. We see that this is not simply knowledge, but unknown and unpredictable knowledge. We want a calculable quantity that fulfills this simple idea.

2.1 SHANNON'S ENTROPY

Claude Shannon did the most to develop what we now call information theory [2, 3, 4]. Although we will begin axiomatically, Shannon motivated his definition differently [2]. We want to determine the average amount of information we receive with a message. This statistical approach has the advantage of being general enough so that it can be applied to anything that has a probability distribution. Following the concept of information described above, a low probability p message has more information than a high probability one. With only this idea, we could create any number of "information" functions ($1/p$, $1-p$, *etc.*). However, using a few well-intentioned axioms, we can narrow down the choice.

Messages, denoted by X, normally have discrete symbols belonging to some alphabet or set $\mathcal{X}$ of finite size A [4][1]. Each individual symbol x taken from this set ($x \in \mathcal{X}$) will then have some probability $p(x)$ determined experimentally by counting the number of occurrences of each symbol in the message[2,3]. We emphasize that this is an experimentally determined $p(x)$ and may differ from the ideal or theoretical probability distribution (assuming such a quantity exists). In general, we will need long messages to determine the probabilities accurately, since our information will be determined by the $p(x)$. (How long is necessary will itself depend on the information as discussed in Sec. 2.3 and Ref. [4].)

We require that our information function[4] of the message, $H(X)$, have the following properties [1, 2]:

[1] A message which has continuous variables is an obvious exception and will be dealt with later.

[2] So if we have a message that is N symbols long and see 'a' N_a times, $p(a) = \frac{N_a}{N}$.

[3] Here we take the pragmatic, "frequentist" approach to the definition of probabilities. Not surprisingly, there is some philosophic controversy surrounding this [7, 44, 45].

[4] $H(X)$ is technically a function of $p(x)$ which is itself a function of x. Thus $H(X)$ is really a functional [4].

1. Similar distributions $p(x)$ should lead to similar values of H. This requirement is mathematically quite sensible.
2. If all the probabilities $p(x)$ are equal, then H must be a monotically increasing function of the number of symbols n. A larger alphabet means more information is transmitted.
3. H must be additive. If we consider the information of two independent messages X and Y, the combined information is $H(X,Y) = H(X) + H(Y)$.
4. We want $H \geq 0$, with $H = 0$ only when a single symbol appears repeatedly[5]. In this case we already know everything from the start.

Taken together, a definition that satisfies all of these criteria is [1]

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = \langle -\log_2 p(x)\rangle_x \tag{2.1}$$

where using the base 2 logarithm gives units of "bits"[6,7]. The quantity $-\log_2 p(x)$ is the information in x which we average over all x, so it is the average information per symbol. When using bits, H is also the average number of "yes" or "no" questions you need to ask to determine the next symbol in a message[8] [2, 3, 4]. From this we can see how H is a measure of unpredictability (uncertainty). If only one symbol appears in the message, we aren't told anything by the message and so $H(X) = 0$. If all symbols are equally likely ($p(x) = 1/A$), each symbol is maximally unpredictable and $H(X) = \log_2(A)$. The connection between unpredictability and information should now be clear: receiving unknown and unguessable knowledge is a measure of unpredictability.

Shannon proposed this definition of information in order to deal with certain problems in communication [2]. There had previously been little success in making rigorous, mathematical progress on the issues of coding and transmission, especially when noise is present. He showed that it is impossible to encode a message such that the code rate (the information per symbol) is less than H. This is the first of his famous theorems, the source coding theorem. An everyday application of this involves compressing the size of data files on a computer. You cannot take a picture or video file on your computer and compress it to a size smaller than its H.

Shannon's second famous theorem, the noisy channel coding theorem, involves the transmission of data through a channel with noise, as its name suggests. Shannon was able to show that one can send a message error-free over any channel even in the presence of noise. This theorem brings other mathematical quantities besides H into play. We will introduce these later.

From this revolutionary beginning in communications, Shannon's entropy has found applications in many other areas such as in biology, where DNA is a hereditary message [41], and in physics, where we now comfortably think of a chaotic system as producing information [17, 42]. Of particular interest is Jayne's

[5] Shannon did not list this as a criterion, but he did discuss it as a feature of his definition.
[6] We will continually express H and other information quantities in bits.
[7] From now on we will not say $x \in \mathcal{X}$ but simply x below the sum, as is usually done [4].
[8] Shannon used this idea to estimate the entropy of the English language [4]

interpretation of statistical mechanics [7]. We maximize the entropy in order to make a least-biased inference of unknown thermodynamic variables. More on this will be discussed later.

There is an anecdotal explanation for Shannon calling H the entropy, a name which we have probably first encountered when learning about heat engines in thermodynamics. Given the resemblance to the statistical mechanical entropy developed by Boltzmann and Gibbs [7, 10], Shannon was supposedly encouraged by Von Neumann to call his information function "entropy" to avoid harsh criticism [38]. According to Von Neumann, since no one really understands the statistical mechanical entropy, Shannon would have an advantage in any debate regarding his H.

2.1.1 Coin toss example

We can discuss some of the properties of H using a simple coin toss experiment. Let the result of the coin toss be called "1" if it is heads and "0" if tails. By flipping the coin many times, we will generate a long stream of 1s and 0s[9]. We might not normally think of this as a message, but we are not treating this string any differently than we would a book or letter[10]. We will thus drop this distinction.

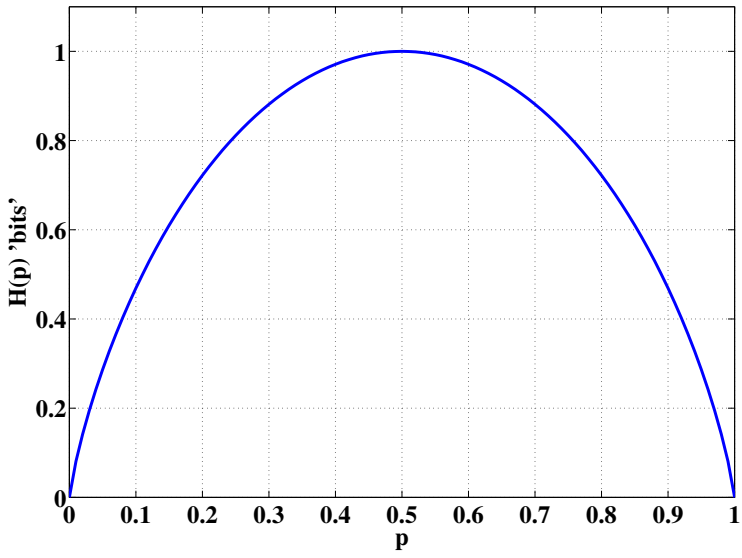


Figure 1: The Shannon information (entropy) H for a coin toss experiment as a function of the probability p that a "1" results. If the coin is fair, we learn a maximum amount of information at each toss since the result is maximally unexpected. If the coin is biased, the result of the toss can be guessed more easily (fewer "yes" and "no" questions on average), and the entropy H decreases.

Suppose the coin is unbiased. The 1s and 0s will then occur with equal probability $p(0) = p(1) = 0.5$.

[9] We will often use strings of 1s and 0s, a binary alphabet, because of the simplicity of presentation.
[10] This also illustrates a deficiency in this definition of information. We can't say if the message has any meaning

The entropy is then easy to calculate as $H(X) = 2 \times (-0.5 \log_2(0.5)) = \log_2(2) = 1$ bits. Each flip provides us with a result that we could not have guessed otherwise.

In the case where the coin is biased, H will have a lower value than this maximum of 1 bit. If the probability for a 1 or 0 is larger than 0.5, we can more easily guess the value of each flip. There is less information (per symbol) gained upon receiving the message. Letting the probability of 1 be p and 0 be $1 - p$, Fig. 1 shows the dependence of H on p for this experiment. This also helps to illustrate the idea of entropy as a measure of unpredictability. Our coin system is maximally unpredictable when $p = 0.5$ but trivially predictable when $p = 0$ or 1.

2.1.2 Literary example

Another example we can use to illustrate the meaning of H is the famous book Moby Dick, by Hermann Melville. Taking the text of this book (which is freely available online [39]), we find that there are slightly over 1 million characters which includes spaces, returns, symbols, *etc.* We reduce the size of our syllabary, which makes the analysis easier and decreases the strain on the statistics. We encode the text such that all symbols and carriages (spaces, etc.) are '0', all numbers are '1' and all the letters are numbered '2' through '27'. We are admittedly losing details, but that will not matter for our present purpose. The probability distribution that we can calculate from doing this is shown in Fig. 2.

From this probability distribution, we calculate that $H \simeq 4$ bits. If the data were completely random, with $p(x) = 1/28$, we would have $H_{max} = \log_2 28 \simeq 4.807$ bits. The value of H thus confirms what we expect and see in Fig. 2: the work of Hermann Melville is not a random collection of symbols. Note that this analysis says nothing of what the information is, only how much there is. Moreover, a strange aspect of this definition of information is that it tells us that a random book would hold more information than most of the books that we read. As long as we keep in mind that this definition is only saying that we learn something new, this is not so strange.

2.1.3 Block entropies and entropy rate

There is a deficiency in this definition of information which was well known to Shannon. This can be illustrated with a periodic string of 1s and 0s: "...001100110011...". Calculating the probabilities of 1 and 0 we find $p(0) = p(1) = 0.5$, just as in the random case. Thus, using Eq. 2.1, we find that $H = 1$ bits. This answer is not wrong, but it goes against the intuition we have developed about entropy. Once we recognize that the message is periodic, the message is not in the least unpredictable. We expect that $H = 0$ bits, but instead we calculated $H = 1$ bits.

We can begin to fix this problem by first considering adjacent pairs of letters and then triplets and so on. By considering the probabilities of blocks of letters instead of just a single letter, we are automatically taking into account relationships between symbols that may extend over some finite "distance" (time, space,

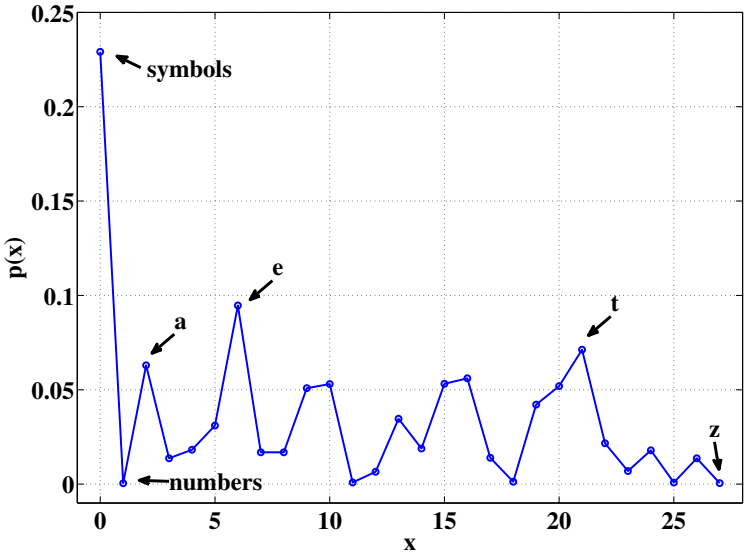


Figure 2: The probability distribution $p(x)$ for the symbols x taken from Moby Dick. Our encoding is that all symbols, spaces, etc. go to $x = 0$, all numbers go to $x = 1$ and all letters are encoded from $x = 2$ to $x = 27$. Clearly some letters are more common than others. A truly random syllabary would have a flat distribution.

etc.) in the message. In other words, we will take correlations into account. A prevalent example of this in the English language is the high probability for the letter q to be followed by u.

Denoting a block of letters of length L as $x^{(L)} = (x_1, x_2, ..., x_L)$, we calculate the $p(x^{(L)})$, which are really joint probabilities[11]. We can then generalize the single letter Shannon entropy to the *block entropies* [4]

$$H(X^{(L)}) = H_L = -\sum_{x^{(L)}} p(x^{(L)}) \log_2 p(x^{(L)}) \tag{2.2}$$

where again the sum is taken over all unique blocks. Considering $L = 2$ as an example, we find that $p(x_1, x_2) = p(x_1)p(x_2)$ if and only if x_1 and x_2 are independent. Thus, $H(X^{(2)}) = H(X_1, X_2) = H(X_1) + H(X_2)$. Since the completely random case necessarily gives the maximum value of the entropy, we see that in general, $H(X_1, X_2) \leq H(X_1) + H(X_2)$ and so also in general, correlations reduce the entropy[12].

Considering our periodic string of 1s and 0s from before, we have for $L = 2$: $p(00) = p(01) = p(10) = p(11) = 0.25$. Thus, $H_2 = 2$ bits. For $L = 3$ we have $p(001) = p(011) = p(110) = p(100) = 0.25$ and so again $H_3 = 2$ bits. The same value is calculated for all $L \geq 2$. Once we took blocks bigger than the period, our block entropies stopped changing in value. All the information is already captured.

[11] The joint probability of two variables is the probability that one symbol *and* the next symbol occur with particular values.
[12] This statement, of course, can be made more rigorous [4].

If we consider a random string, we find that $H_L = L$. As we consider larger blocks, our uncertainty increases accordingly. We have managed to distinguish between a random and a periodic string, and we can also in principle incorporate all inter-symbol dependencies. However, we have no way of choosing which L to use.

This problem leads us to define an *entropy rate*[13] h which will be independent of L by construction. The amount of information contained in blocks of size L is of course H_L. The amount of information in H_{L+1} is all that contained in H_L plus one more symbol, with all of the inter-symbol dependencies up to a length L taken into account. The quantity

$$h_L = H_{L+1} - H_L$$

is thus the information per symbol with interdependencies up to a distance L taken into account. If correlations in the system are finite in extent, then there is a well-defined limit such that [4]

$$h = \lim_{L\to\infty} h_L = \lim_{L\to\infty} H_{L+1} - H_L \tag{2.3}$$

This is equivalent to the definition $h = \lim_{L\to\infty} H_L/L$ as well as another that will be presented later, if the process is stationary [26, 4]. However, our first definition here is the best in the sense that it approaches a limit fastest when it exists [26]. The entropy rate h is 0 for periodic messages and 1 for random ones (if using a binary alphabet with base 2 logarithm). This h is the information revealed by the next symbol in a message, given knowledge of all the preceding ones. This is our ultimate measure of unpredictability. Even knowing the infinite past, we may not know exactly what will happen next.

2.1.4 Compression

An alternative method for determining h is based on data compression. The length of a message S_0 is the total number of symbols in the message. For an ASCII text file on a computer, this would be the total number of ASCII characters (plus any headers and other syntax). The lower limit for the length to which a message can be compressed S from its original length S_0, for any compression algorithm, is its entropy: $S \geq H$ [4, 26, 109, 110]. As mentioned earlier, this is just an application of Shannon's first theorem.

Let us look at how compression works. Every message is written in some kind of code or alphabet. Morse code was designed to shorten the overall length of messages by assigning more frequent letters to shorter symbols (*e.g.* *e* is a single dot). We could design a more complicated code which looked at pairs and triplets of letters. For example, we might be able to shorten the overall length of the message by assigning short symbols to frequently used words like *a*, *an* and *the*. This would quickly become cumbersome for any telegraph operator, but it's no trouble at all for computers. Moreover, they are not confined to a single code book but can write their own new code book. This is in fact, what a compression algorithm does: it re-expresses the message in a shorter form.

[13] When applied to physical data sets, the convention seems to be that for a time series the word "rate" is used while for spatially extended series, "density" is used instead.

It is fascinating that the length of this rewritten message is limited by it information. Compression provides a nice way of thinking about how much information is contained in a message, since it reduces the message to its irreducible "essentials". There is no way to shorten a completely random message since each symbol is independent of the others. All symbols and words are of equal probability, so no rewriting will improve on the message's length. For a repetitive stream of symbols (like "...111111...") the message is trivially compressed to almost zero size.

We define the *compression ratio* c as [109]

$$c = \lim_{S_0 \to \infty} \left(\frac{S}{S_0} \right) \times \log_2 A, \tag{2.4}$$

where A is the alphabet size (*e.g.* for a binary alphabet $A = 2$). We must use compression schemes based on the Lempel-Ziv algorithm, which is optimal in the sense that c converges to h in the limit of infinite S_0 [4, 109]. The value of c is independent of file type but does require that the compression program be based on this Lempel-Ziv algorithm. In practice, we must account for a file's "overhead" (file headers, *etc.*). A random data set is compressed and that compression ratio is used to normalize the real data [111].

Traditionally, c has been given the name of "algorithmic" or "Kolmogorov complexity" [4, 109, 111]. This is a measure of the computational complexity of the data set in question. That is, what is the shortest computer program that can reproduce this data set? Even if c is not equal to h due to a finite S_0, it is still a measure of the information content of the data [111, 112, 113]. It is important to recognize the many limitations involved in calculating information content, as will be discussed in Section 2.3.

For fun, we have included a plot of the compression ratio for several well known books in Fig. 3. It is surprising that a famous philosophy manuscript has a smaller amount of information than some children's books. This is probably because in the sophisticated arguments of Kant, there is order and repetition. Also, every one knows that good children's books are especially exciting and unpredictable, so it is not really a fair comparison. This example serves to remind us that our technical definition of "information" here is very different from the usual sense of the word. This difference must be kept in mind.

2.1.5 Other information quantities

Several other information quantities will be useful in our discussion. They are variations on the theme of how much information we gain, on average, from receiving a message. We may also want to ask about the information or unpredictability of two or more messages. How does getting one message affect the unpredictability of another? What do they share?

2.1.5.1 Joint entropy The first of these quantities has already been presented above in another form. The *joint entropy* $H(X, Y)$ of two messages (or variables) X and Y is the information or uncertainty contained in both X and Y [4]. These may be two messages or a record of two physical variables such as temperature and pressure. Using the joint probability distribution $p(x, y)$, which is the probability of both x and y

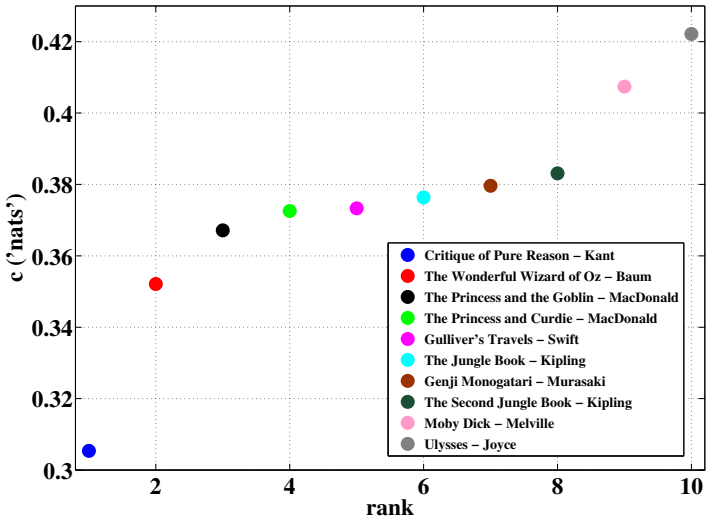


Figure 3: The compression ratio for several well known books.

occurring, we have

$$H(X,Y) = -\sum_{x,y} p(x,y)\log_2 p(x,y) \tag{2.5}$$

again with the sum over all x and y. $H(X,Y)$ tells us how unpredictable the messages X and Y are when received together. As mentioned in the previous section, this quantity satisfies the relation $H(X,Y) \leq H(X) + H(Y)$, with equality if X and Y are statistically independent. In other words, if X and Y share information about each other, then receiving both messages generally gives us less information than when we consider them apart and add up the information. One may of course consider the joint entropy of many variables.

2.1.5.2 Conditional entropy Another quantity of interest is the *conditional entropy*. Whereas the joint entropy $H(X,Y)$ gives the information in both X and Y, our conditional entropy $H(X|Y)$ gives us the information we receive from X given that we already know Y [4]. Or, how unpredictable is X given that we know Y. This requires calculating the conditional probabilities $p(x|y)$, which of course means the probability of x given y. As an example, one could ask what is the probability that a person has long hair, given that they are a woman.[14] The conditional entropy is defined as

$$H(X|Y) = -\sum_{x,y} p(x,y)\log_2 p(x|y). \tag{2.6}$$

[14]Using Baye's theorem $P(X|Y) = P(X)P(Y|X)/P(Y)$ and some appropriate guesses, we could turn this around to ask what is the probability that a person is a woman, given that they have long hair.

The reason for the use of the joint probability $p(x, y)$ in this definition is that we needed to average our information about X given Y over all values of Y, which changed the conditional probability into a joint one through $p(x, y) = p(y)p(x|y)$. Explicitly, we have

$$H(X|Y) = \langle H(X|Y = y)\rangle_y = \sum_y p(y)[-\sum_x p(x|y)\log_2 p(x|y)],$$

which leads to the above definition in Eq. 2.6. Moreover, if X and Y are independent, then $p(x, y) = p(x)p(y)$ and $p(x|y) = p(x)$, so $H(X|Y) = H(X)$. That is, Y tells us nothing about X, so we gain the full amount of information from viewing X without any help or input from Y.

Clearly the joint entropy and conditional entropies are related, since we are talking about the information gained from two messages or variables. Suppose we received message Y first. We gain an amount of information $H(Y)$ from this message. If we then receive the message X, we gain an amount of information $H(X|Y)$ from this message, since Y may have some shared information with X. However, the information contained in both messages is the joint entropy $H(X, Y)$, which tells us that

$$H(X, Y) = H(Y) + H(X|Y) \tag{2.7}$$

as expected. This can be shown formally by manipulating the definition of $H(X, Y)$ and using that $p(x, y) = p(y)p(x|y)$ [4].

We can illustrate these new entropic quantities with our coin flip experiment and Moby Dick. In the case of Moby Dick, we have already calculated $H(X) \simeq 4$. If we let Y be the previous letter in the story before X, then we find that $H(X, Y) \simeq 7.32$ bits and $H(X|Y) \simeq 3.32$, which confirms the relation $H(X, Y) = H(Y) + H(X|Y)$. Apparently, knowing the previous letter does in fact reduce our uncertainty about the next letter, since there are of course correlations in a good book. As for the coin toss experiment, we can again consider a biased coin. Letting Y be the toss prior to X, we would find that $H(X|Y) = H(X)$, since they are not correlated at all even in a biased coin toss. Thus $H(X, Y) = H(X) + H(Y) = 2H(X) = 2$ bits. It will take twice as many yes and no questions to guess the correct outcome of two consecutive coin tosses.

We can use the conditional entropy to define the entropy rate in another way and gain a slightly different perspective. Let $X = X_i$ and $Y = X_{i-1}, X_{i-2}, ..., X_{i-L+1}$. That is, we take strings of length L and calculate the entropy of the last symbol X_i conditioned on the previous $L - 1$ symbols. The entropy rate is thus

$$h = \lim_{L\to\infty} H(X_i|X_{i-1}, X_{i-2}, ..., X_{i-L+1}) \tag{2.8}$$

where, of course, the index i is not important but simply a place holder. So now we see that the entropy rate h is really the information we gain from receiving a symbol, given knowledge of all previous symbols.

2.1.5.3 Kullback-Leibler divergence The next quantity we consider is similar to the conditional entropy in that it tells us about the relationship between two things. The *Kullback-Leibler divergence* or *relative*

entropy, $D(p||q)$, is a measure of the difference between two distributions p and q for the same variable or message X [4]. It is not symmetric in p and q and so can't be considered a proper "distance" function[15], but it turns about to be quite useful and has a wide variety of applications. We will use this later to quantify intermittency. We define

$$D(p||q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \tag{2.9}$$

and one can easily see the lack of symmetry in the definition. We also see that if $p(x) = q(x)$, then $D(p||q) = 0$. We can see how $D(p||q)$ tells us about the difference between two distributions by noting that if we have some "real" distribution $p(x)$ and we compare it with the random, equidistribution $q(x)$, then we have

$$D(p||q) = -\sum_x p(x) \log_2 q(x) + \sum_x p(x) \log_2 p(x) = H_{max} - H[p(x)] \tag{2.10}$$

so this tells us the difference in the amount of information contained in our message from that of a truly random message. That is, this is in some sense how much we do know about our system (if we equate the use of the equidistribution with knowing nothing).

One interpretation of $D(p||q)$ is that if we knew the true distribution of a message or data $p(x)$, then we could compress the data to a size $H(p)$[16]. If we find ourselves in a circumstance where we don't know the true distribution $p(x)$, we may guess at some distribution $q(x)$ and use that to encode our message[17] and compress it. We are able to compress our message to a size $H(p) + D(q||p)$ in this way.

Suppose, for example, that we wanted to represent a biased coin flip with an unbiased one. Then $p(1) = p$ and $p(0) = 1 - p$, but $q(0) = q(1) = 0.5$. Thus $D(p||q) = p \log \frac{p}{0.5} + (1-p) \log \frac{1-p}{0.5}$ and $H(p) = -p \log p - (1-p) \log(1-p)$. For any p, $H + D = H_{max} = 1$, even though D grows as p departs from 0.5. We could have compressed our message to $H(p)$, but with our horrible choice of q we are not able to compress at all.

If instead we have $p = 0.3$ but make the less horrible mistake of thinking that $p = 0.4$, we calculate $H + D \simeq 0.9125$ bits with $H(p) \simeq 0.8813$ bits and $D \simeq 0.0312$ bits, respectively. So of course we didn't do as well as if we had known $p = 0.3$, but we didn't do as badly as before.

2.1.5.4 Mutual information We can construct a quantity using $D(q||p)$ that is symmetric with respect to the two messages X and Y, even though $D(q||p)$ is nonsymmetric in p and q. We let $p = p(x, y)$ and $q = p(x)p(y)$. This quantity is called the *mutual information* between X and Y:

$$I(X;Y) = D(p(x,y)||p(x)p(y)) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2.11}$$

and is in some sense a measure of the difference between an independent joint probability distribution $p(x)p(y)$ and the real joint probability distribution $p(x, y)$. If X and Y are independent, then $I(X;Y) = 0$,

[15] To fix this lack of symmetry in its arguments, people sometimes use $\frac{1}{2}D(p||q) + \frac{1}{2}D(q||p)$ instead.

[16] We write the entropy in this way to highlight its dependence on the distribution

[17] Encoding a message to compress it is intimately related to the entropy by Shannon's theorems. Clearly if a particular symbol occurs more frequently than another, we will choose a shorter description of that more frequent one to reduce our message's length overall.

otherwise, this is a measure of how much information they share. The idea of sharing information is nicely represented by the diagram below in Fig. 4.

This can be written in another way in terms of conditional entropies as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

which expresses mathematically what we have already said with words. If X and Y have nothing to do with each other, then $H(X|Y) = H(X)$ and so $I(X;Y) = 0$. If X and Y are in fact related, then they share an amount of information equal to the full amount of information in X, $H(X)$, less the amount that this information is reduced to if conditioned on Y, $H(X|Y)$.

Shannon used the mutual information to express his second theorem. All channels (telephone, telegraph, smoke signal, *etc.*) use a set of symbols to convey a message. Let us call this message X. We can use the symbols however we want, which means that we can assign any distribution $p(x)$. There is some best way which achieves the highest possible rate of transfer of information $T = \sup_{p(x)} I(X,Y)$, where Y is the received message. Y might be different from X since there will always be some errors due to noise that contaminate the transmission of the message. Shannon showed that if you send your message at a transmission rate $R < T$, then you can always express it in such a way that the errors are arbitrarily small.

We can visualize the relationships between some of these information quantities using an "information diagram" as in Fig. 4. Consider two messages X and Y. Each have an amount of information $H(X)$ and $H(Y)$. They also share an amount of information about each other $I(X;Y)$ shown by an overlap of the two ellipses in Fig. 4. The union of the two ellipses is the joint information $H(X,Y)$, the total amount of information we get from receiving both messages. The relative complement of each (the part of $H(X)$ that doesn't intersect with $H(Y)$) is clearly not shared, and so represents the amount of information we get from the message X already knowing message Y, *i.e.* $H(X|Y)$.

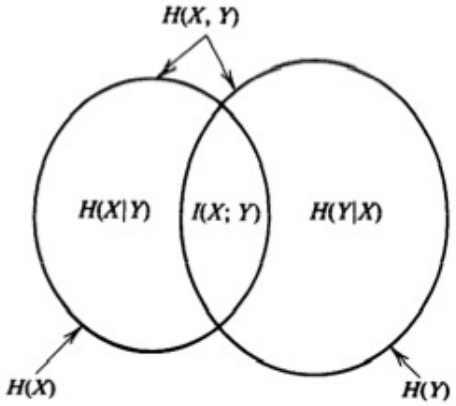


Figure 4: The relationships between the different information quantities can be visualized with an "information diagram" [4].

2.2 EXCESS ENTROPY

We have developed h to characterize messages, but we will need more tools than this. When using h, recall that it achieves its extreme values for total randomness: $h = 1$ bits, or complete (and perhaps boring) order: $h = 0$ bits. Most of the systems we will deal with are neither. Another deficiency in h is that it fails to distinguish between periodic strings with different periods T_1 and T_2. Both have $h = 0$, but even a quick glance at either message will show that they differ. Moreover, h is a measure of unpredictability, but we would also like to say something about what and how we can predict.

The mutual information can be used to define a quantity E that is able to distinguish between our two periodic strings. The *excess entropy* E is the mutual information between the (semi-)infinite past and (semi-)infinite future of a message. The present is taken anywhere in the message. It answers the question: how much do the next symbols have to do with the previous ones? It is a memory. Defining (with arbitrary) i, the past $\overleftarrow{X} = (...X_{i-3}, X_{i-2}, X_{i-1})$ and the future $\overrightarrow{X} = (X_i, X_{i+1}, X_{i+2}, ...)$, we can define E mathematically as

$$E = I(\overleftarrow{X}; \overrightarrow{X}) = H(\overrightarrow{X}) - H(\overrightarrow{X}|\overleftarrow{X}) \tag{2.12}$$

This mathematical expression is hopefully less vague than our description of E above. This is the reduction in unpredictability of the future from knowing the past.

Also, like h, $E \geq 0$ with $E = 0$ for only two cases (that we know of). If a message is random with $h = 1$, then by definition there are no correlations, $H(\overrightarrow{X}|\overleftarrow{X}) = H(\overrightarrow{X})$, and so $E = H(\overrightarrow{X}) - H(\overrightarrow{X}|\overleftarrow{X}) = 0$. If we have a message which is just an infinite repetition of a single symbol[18], we likewise have $E = 0$. This is because we have no uncertainty about the future or the past, and no information is gained from either, so $H(\overrightarrow{X}) = H(\overrightarrow{X}|\overleftarrow{X}) = 0$ and so $E = 0$. Knowing the past didn't reduce our unpredictability of the future at all.

As with h, there are several other ways of calculating E. The above definition is perhaps the easiest to understand but not the easiest to calculate. Instead, we calculate E from the approach of the h_L's to h.

$$E = \sum_{L=0}^{\infty}(h_L - h) = \sum_{L=0}^{\infty}((H_{L+1} - H_L) - h). \tag{2.13}$$

The value of E depends on the speed and degree of the approach. For the random case, $h_1 = h$ and so $E = 0$. In practice we find a finite $L = L^*$ where $h_{L^*} = h$, in which case we can simply write

$$E = (H_{L^*+1} - H_0) - (H_{L^*+1} - H_{L^*}) = (H_{L^*} - H_0) = H_{L^*}$$

since $H_0 = 0$. This is the reason for calling E the excess entropy. We needed to know an extra amount of information H_{L^*} in order to determine the true amount of information (per symbol) in the message, which is h. This can be understood graphically as well, where E is the "area" beneath the curve $h_L - h$ as in Fig. 5.

[18] This can be thought of as a periodic string of period 0.

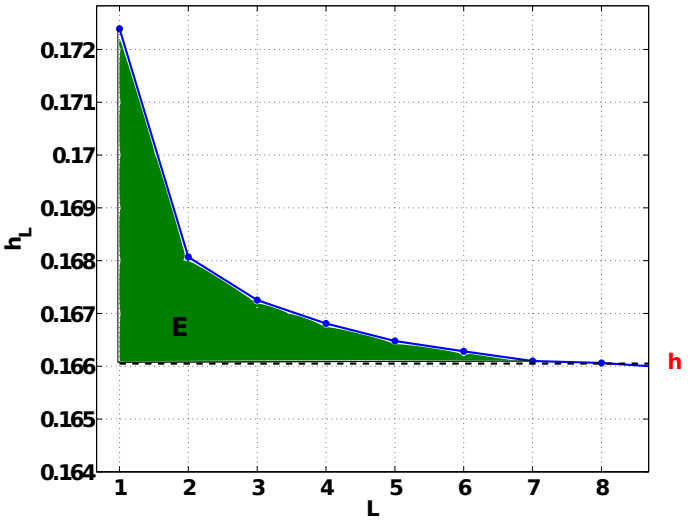


Figure 5: The excess entropy E shown graphically as a measure of the approach of h_L to h. The faster the approach, the smaller is E.

Returning to our two periodic strings with period T_1 and T_2, we see that $L^* = T_{1(2)}$, since it takes a block this long to fully incorporate all of the behavior of the string. We can also see that there are $T_{1(2)}$ unique blocks with equal probability $p = \frac{1}{T_{1(2)}}$ and so

$$E = H_{L^*} = H_{T_{1(2)}} = T_{1(2)} \frac{1}{T_{1(2)}} \log_2 \frac{1}{T_{1(2)}} = \log_2 T_{1(2)}.$$

So E is the logarithm of the period for periodic strings. For example, consider the string ...010101... and ...011011011... with periods $T = 2$, 3 respectively. Here $E = \log_2 2 = 1$ bits and $E = \log_2 3$ bits respectively.

There are many interpretations of E that depend on the context of its use. An idea that we will latch on to here is its role in prediction. Because E contains all the information about the future we get from the past, any prediction method must *at least* use this much information. This doesn't tell us how to do it, but places a lower bound on whatever method we do use [21, 22].

2.3 DIFFICULTIES IN ESTIMATING ENTROPIES

We will take a small detour here to deal with the actual computation of the quantities h and E from real data. Since determining E largely depends on determing h, we will focuse on h. The plot in Fig. 5 is deceptively simple. We kept the behavior of h_L as L increases hidden because it takes a turn for the worse as shown in Fig. 6. The strong downward trend in h_L as L increases is due to the finiteness of the data. If the message

is not long enough, then the probabilities of the words can not be determined accurately. The message will look increasingly non-random, since certain words will appear more than others by "accident". Even if we took data from a truly random source, we won't realize this if we don't have enough equal instances of all the unique words. In the extreme limit of $L \to \infty$, there is only one word with probability unity, the entire message, and so $h = 0$ by definition.

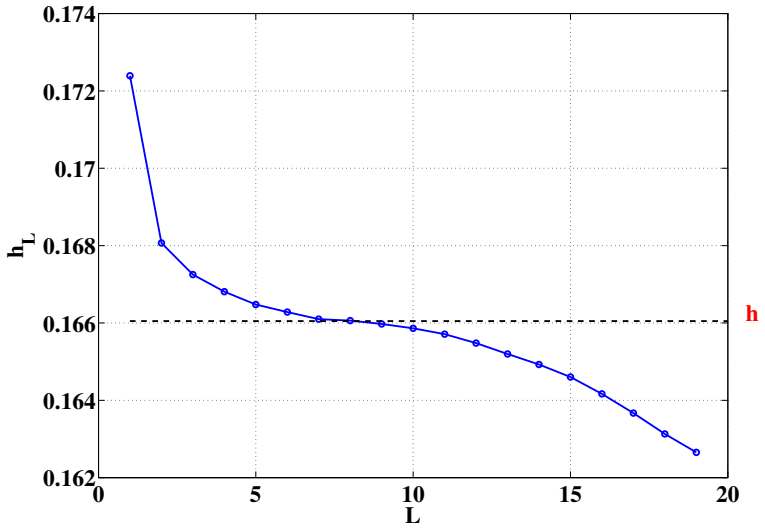


Figure 6: The excess entropy E shown graphically as a measure of the approach of h_L to h. The faster the approach, the smaller is E.

So how do we know how far in L we can calculate h_L before we can no longer trust the result? There is no perfect answer to this question. One can try to estimate the maximum L that can be trusted by assuming the Asymptotic Equipartition Property (AEP) and the idea of the "typical set". Without going into details, this is simply a formal way of expressing the idea that "...almost all events are almost equally surprising..." [4]. Of course, this equiprobability doesn't mean that the events are random with a maximum h. Rather, we are saying that mathematically

$$-\frac{1}{L}\log_2 p(X^{(L)}) \to h$$

in the limit of large L. At the same time we presume that most words[19] $X^{(L)}$ are equiprobable with $p(X^{(L)}) \simeq 1/N(X^{(L)})$, where $N(X^{(L)})$ is the number of unique words of length L. Then $N(X^{(L)}) \simeq 2^{hL}$. Of course for the random case $h = 1$ this is necessarily true. We then say that the largest L that we can safely estimate h_L for is $L_{max} \simeq \frac{\log_2 N}{h}$, where N is the size of our data set (length of our measurement). Since we must use the calculated h to determine the maximum L, this essentially serves as a self-consistency check. If h was determined at an L larger than we can safely estimate, then we have obviously done something wrong.

[19] In the so-called "typical set" [4].

There are also several fancy methods for estimating the bias due to the finiteness of the data [26]. All of these suffer from making overly strong assumptions. After all, we are trying to estimate the penalty of our ignorance without knowing what is really going on. In the end, it appears that the estimates for the bias do not really differ all that much from the untampered estimate [18]. In practice, we will simply determine h by finding where the h_L's appear to approach a constant value and before the downward trend. In other words, we will let h be the value of h_L at its inflexion point. Looking back at Fig. 6, it is encouraging that all the points surrounding the inflexion at $L = 8$ have almost the same value. Our estimate of h is not sensitive to our ability to find the inflexion. There is a considerable body of literature on entropy estimation (see for example Ref. [26] and included references), but many require uncomfortable additional assumptions. Moreover, we did not find that the results change significantly.

2.4 INFORMATION THEORY APPLIED TO PHYSICAL SYSTEMS

It is perfectly sensible at this stage to challenge the use of information theory with physical systems. After all, this is a set of tools originally developed for analyzing messages sent between people. What can we hope to learn from applying this to physical systems? We have already given the example of the coin flip, but perhaps this is a specially contrived system. Since information theory has been and is being used fruitfully for the study of many physical systems, let us look at several examples to convince ourselves that we are not being foolish. Keep in mind, however, that while it is the same information theory in all cases, the purpose and philosophy are quite different.

2.4.1 Shaw and the information of chaotic systems

Chaos may be loosely defined as that inherently unpredictable behavior that arises in nonlinear physical systems. This is characterized by a strong sensitivity to initial conditions. Despite starting the same system with nearly the same beginning, the subsequent behavior is completely different[20].

One of the most useful tools for studying the behavior of dynamical systems in general and chaotic systems in particular is the Lyapunov exponent(s) [47]. These are most easily understood by imagining the evolution of trajectories in phase space, a caricature of which is displayed in Fig. 7. Let $x^i(t)$ be the time-dependent vector of the system's coordinates in phase space[21] Consider two initial conditions $x^i_A(t=0) = x^i_{A,0}$ and $x^i_B(t=0) = x^i_{B,0}$ separated initially by $\delta x^i_0 = |x^i_{A,0} - x^i_{B,0}|$. For a chaotic system, $\delta x^i(t) \simeq \delta x^i_0 e^{\lambda_i t}$, with positive λ_i, the Lyapunov exponents indicating exponential separation of trajectories[22]. Of course, both

[20] This is practically important for the prediction and modeling of systems such as the weather, where unavoidable rounding errors can lead to vastly different predictions. Incidentally, the forecast "percentage" is based on the number of trials out of the total that give that particular result [46].

[21] For a 3D N-particle system, this would be a $2 \times 3 \times N$ dimensional vector of the positions and momenta of each particle.

[22] This was perhaps first appreciated by Ed. Lorenz when he simulated a simplified version of the weather on his (now) primitive computer [48].

will stay in the allowed region of phase space, so this exponential behavior will not go on forever. When their separation approaches the size of the allowed region of phase space then the confinement suppresses the exponential behavior. (Also, depending on the character of the system there may still be $\lambda_i \leq 0$[23].)

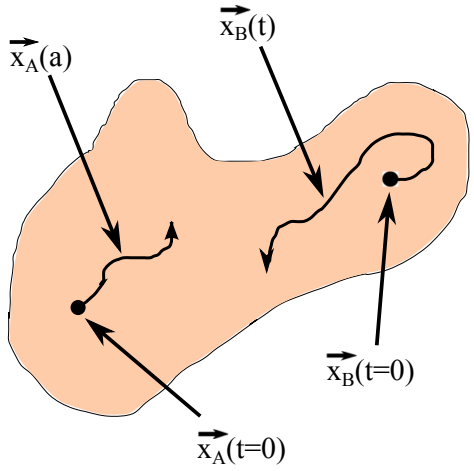


Figure 7: A caricature of phase space. Two trajectories are shown in the colored space, which represents the allowed region determined by physical law possibly expressible in mathematical form. The two trajectores begin at different points at time $t = 0$ and evolve in time afterwards.

At the beginning of the so-called chaos revolution, R. Shaw wrote an important paper entitled "Strange attractors, chaotic behavior, and information flow" [17]. In this paper, Shaw made the following connection between the information entropy defined by Shannon and the behavior of dynamical systems. As two trajectories separate in phase space ($\lambda > 0$), they are creating new information. This is because their separation can be described with more and more detail. Alternatively, we might say that we need more information to tell a friend about how they differ as time progresses.

Consider the two exponentially separating lines in Fig. 8. At a finite resolution (as will always be the case with real experiments and even simulations[24]) we can only express the difference between the trajectories in terms of the number of sections (see Fig. 8) between them at any point in time. In these units $\delta x(t_A) = 1$ section. Written in binary code, this becomes 0. Likewise $\delta x(t_B) = 5$ sections $= 101$ and $\delta x(t_C) = 18$ sections $= 10010$. The description or message is getting longer[25]. This simple idea has a much more rigorous expression in Pesin's theorem, which for most chaotic systems equates h[26] and the sum of all positive Lyapunov exponents [42, 18].

[23] If this system is dissipative then $\sum \lambda_i < 0$, since the total volume of phase space is shrinking. There may also be oscillations in other degrees of freedom.

[24] Actually, nature itself operates at finite resolution. This graininess is expressed by Heisenberg's uncertainty principle [17, 10]

[25] Don't worry that we are looking at a single trajectory even though the entropy is a statistical quantity. The Lyapunov exponents are experimentally determined by averaging over many trajectories.

[26] Actually it is the Kolmogorov-Sinai entropy h_{KS}, which after chopping the phase space into regions of size ϵ is $\sup_\epsilon(-\sum p_i \log p_i)$, where p_i is the probability of finding a trajectory in the ith region [42].

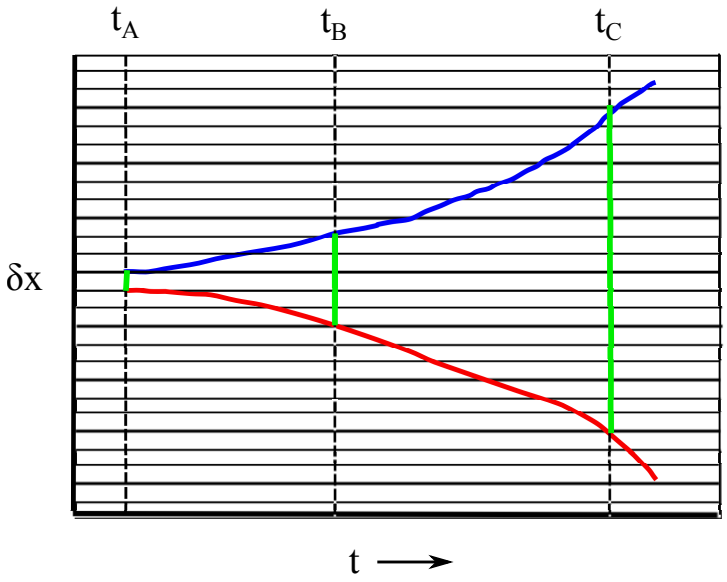


Figure 8: A caricature of two trajectories separating in phase space. The phase space is split up into small sections ("partitioned") corresponding to the resolution of the measurement. As time progresses, more sections lie in between the two trajectories.

2.4.2 Jaynes and statistical mechanics

It is worth noting as E. T. Jaynes does, that [7]

> ...the mere fact that the same mathematical expression $-\sum p_i \log p_i$ occurs both in statistical mechanics and in information theory does not in itself establish any connection between these fields. This can be done only by finding new viewpoints from which thermodynamic entropy and information theory entropy appear as the same concept.

In this same paper, and elsewhere, he attempts to make this connection.

Consider a system with some quantity x that can take on discrete values 1, 2,...n. Given only the expectation value of some function $f(x)$[27]

$$\langle f(x) \rangle = \sum_i^n p_i f(x_i),$$

we want to determine the expectation value of some other function $g(x)$. This is a typical statistical inference problem and is easily "solved" by maximizing the entropy of the yet to be found distribution $H(x) = -\sum p_i \log p_i$ subject to the constraints $\sum p_i = 1$ and $\sum p_i f(x_i) = \langle f(x) \rangle$.

The reason for maximizing H[28] from the point of view of information theory is that in order to make the least biased guess possible, we must maximize our uncertainty. Any lower value of H would mean that we

[27]The expectation value is the average value of that quantity which is necessarily determined by the probability distribution of its independent variables.

[28]We don't need to push forward and calculate h since the x's are independent. In this case we already have $H = h$.

added some additional constraint to the problem that we don't have or don't know about.

This is the same procedure we use in statistical mechanics to find, for example, the canonical ensemble probability distribution. (In that case $\langle f(x) \rangle$ is the average energy of the system.) Now, however, we have a motivating reason for doing this. In Jayne's view, we should interpret statistical mechanics as a theory of statistical inference. We then have no need for assumptions such as ergodicity, *etc.* For him and others, this makes the whole game more pallatable. He also champions the extension to nonequilibrium statistical mechanics, where he has apparently met with some success [7, 8, 9].

2.4.3 Landauer and physical law

While we seem to be focused on what information theory has to tell us about physical systems, it is really a two-way street. All real messages are sent through some physical channel, and the laws of physics impose restrictions or limits on this process. Since computers are really just communication channels moving information to and fro, this has especially strong implications for computer science. Landauer and others have worked out many of the physical limits of computation[29], including making some of the first steps into quantum computation [147, 15].

Knowing that computation is restricted by physical laws, Landauer and others have gone further to suggest that nature itself can be thought of as a computer moving and transforming information. Of course, this computation will obey the physical laws when properly understood. This has led to one particular resolution of the paradox involving Maxwell's demon. Once information is given physical footing, the decrease in thermodynamic entropy achieved by the demon is exactly compensated by an increase in information. This interpretation has been hotly contested by some number of people [127, 128]. Whether this is a correct view of reality or not is left to the reader to decide for themself.

2.5 LOGISTIC MAP

Let us put Shaw's ideas and our information tools to practice. We will use the logistic map, a simple one-dimensional nonlinear map which nicely illustrates chaotic behavior [106]. It also has considerable historic significance from the chaos revolution [106]. The value of a variable x is determined from its previous value according to the rule:

$$x_{n+1} = r x_n (1 - x_n) \tag{2.14}$$

where r is a parameter that increases the strength of the nonlinearity. As r increases, the system goes through a series of period-doubling bifurcations[30] and eventually becomes chaotic at $r \simeq 3.56995$ ($\lambda > 0$). As usual, $x \in (0, 1)$ and $r \in [0, 4]$. As mentioned earlier, Pesin's theorem states that the sum of the positive

[29] Not surprisingly, there are also connections with Maxwell's demon, a kind of devilish computer in its own right.
[30] This means that a fixed point will split into two other fixed points.

Lyapunov exponents λ is equal to h, as long as the system satisfies certain conditions [42]. For the logistic map, which is one-dimensional, there is only one λ for each value of r. The value of λ has been calculated as a function of r, using the algorithm in [114]. We also estimate the entropy rate h as described in Subsection 2.1.3 and Section 2.3. We include the compression ratio c from Section 2.1.4 as an independent estimate of h. These are all plotted together in Fig. 9. For each value of r, a randomly chosen initial condition is iterated 10^6 times.

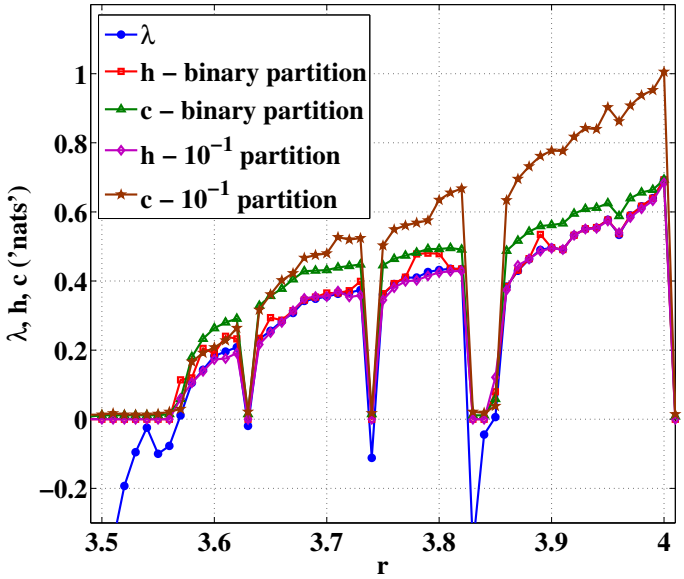


Figure 9: Lyapunov exponents λ ($\bullet$) for the logistic map plotted as a function of r (see text). The entropy rate h (binary: $\square$, 10^{-1}: $\diamond$) and compression c (binary: $\triangle$, 10^{-1}: $\star$) are plotted for two different partitions. Although c does not follow h and λ perfectly, it still follows the same trend.

The results from using two partitions[31] are shown in Fig. 9. A binary partition is used where $x = 1$ if $x \geq 0.5$ and $x = 0$ if $x < 0.5$, so $\epsilon = 0.5$. (The location of this partition divider is important [115].) The second partitioning involves simply rounding the data to the first decimal point ($\epsilon = 10^{-1}$) and assigning a symbol to each distinct data value. The estimate h performs very well, while c shows significant deviations for the 10^{-1} partition. Despite its shortcomings in estimating h, c is nonetheless useful as a measure of the information contained in these finite sequences. It follows the same trend as λ and reveals the logistic map's information dependence on r. The values of r for which λ is negative have $h \simeq 0$, since it is positive definite.

Partitions with as few as 2 slices or as many as 1000 slices give essentially the same h and c. This is because the partitions are all generating, *i.e.* they represent the dynamics faithfully and the entropy calculated for any of them is the Kolmogorov-Sinai entropy h_{KS} [26]. In other words, anything finer than a

[31] A full discussion of partitions is left until Sec. 4.6.1.2. For now, think of it as the dividing up of phase space as depicted in Fig. 8.

binary partition is overkill. This isn't always true but suggests that crude representations of data can still capture important features. This emboldens us to do the same for turbulence, to be discussed in Section 4.6.

Once the transition to chaos occurs, λ and the estimates of h increase almost monotonically. There are several isolated regions where the logistic map returns to periodic behavior [106] and so $\lambda < 0$ and $h \simeq 0$. The general behavior appears to be that as the strength of the nonlinearity increases (see Fig. 9), so does h. Chaos creates information. Similar behavior was observed at the onset of turbulence in Taylor-Couette flow [102]. This increase in h for the logistic map is accompanied by a decrease in the strength of correlations, as will be shown shortly.

2.5.1 Modified Logistic Map

In order to get a better picture of the importance of correlations, a modified logistic map is introduced to explicitly increase correlations through a term that couples to previous iteration values further back than one. The modified logistic map is defined as

$$x_{n+1} = f(x_n) + d\Big[\frac{f(x_{n-2}) + f(x_{n-1})}{2} - f(x_n)\Big] \tag{2.15}$$

where $f(x) = rx(1-x)$ and d is the coupling strength. This modification is really a kind of logistic delay map [106]. The logistic map had each step directly coupled to one previous step. This adjustment here couples each next step to the previous three steps. Now using three random initial conditions, this map is also iterated 10^6 times and the compression estimate c is used to compare h for different values of d. The results are shown in Fig. 10.

Even for small d, c is changed drastically. As d is increased, c decreases and the transition to chaos shifts to larger values of r. In order to quantify correlations for messages, we can use the mutual information introduced earlier in Subsection 2.1.5 (see Eq. 2.11): $I(X;Y)$. When the two variables X and Y are symbols separated by a certain number of symbols Δ ($X(i)$, $Y = X(i+\Delta)$), then $I(\Delta)$ becomes like an autocorrelation function for symbolic sequences (called "transinformation") [116, 141]. We calculate the decay rate of this transinformation

$$\gamma_C = 1/\sum_{0}^{\infty} I(\Delta) \tag{2.16}$$

as a means of comparing the strength of correlations for different values of r and d.

The mutual information is observed to decay exponentially for more than a decade in Δ for the chaotic regime of the logistic map and the logistic delay map with a decay rate γ_C that increases with r (see Fig. 11). Put another way $1/\gamma_C$, which can be thought of as a correlation time, decreases with r. The correlations are thus decreasing as the strength of the nonlinearity increases, which corresponds well with the understanding that h is reduced by correlations. Figure 11 shows γ_C as a function of r for three different values of d. The addition of coupling has increased the strength of the correlations, and mirrors the drop in c.

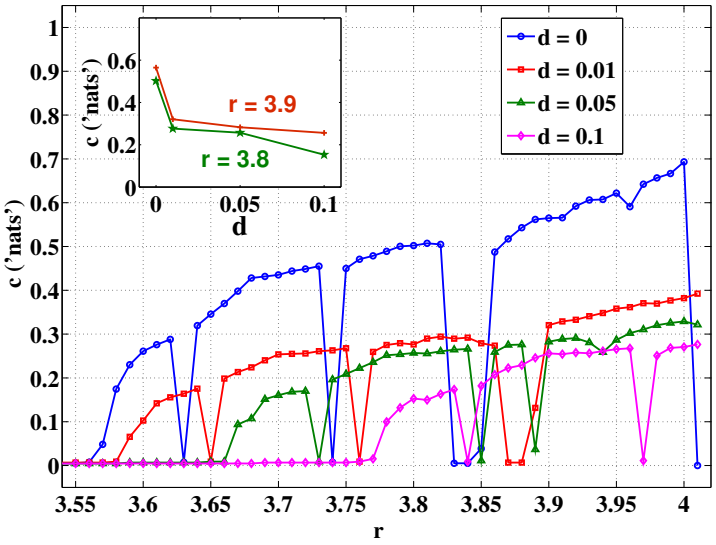


Figure 10: Entropy rate estimate c for the modified logistic map using a binary partition. (Recall that a binary partition is generating for the logistic map.) As d increases, c is lowered considerably. ($d = 0$: ◯, $d = 0.01$: □, $d = 0.05$: △, $d = 0.1$: ⋄.) Inset: c vs. d for fixed r in the chaotic regime.

2.5.2 Excess Entropy of the Logistic Map

While we have already described the general behavior of the logistic map (modified and regular) using h with the mutual information I, we could instead pair h and E. After all, E is defined as the mutual information between the (semi-)infinite past and (semi-)infinite future. We used the decay rate γ of I above, but in fact we can show that E is related to the decay rate γ. This stems from an *empirical* observation that for most chaotic systems [27] we find that

$$h_L - h \propto 2^{-\gamma L}. \tag{2.17}$$

Letting the constant of proportionality be B, we see that we can write this as a simple geometric series as long as $2^{-\gamma} < 1$. Also, since for $L = 1$ we have $h_1 - h = H_1 - h = B$ we can rewrite this as

$$E = \sum_{L=0}^{\infty} B2^{-\gamma L} = B \sum_{L=0}^{\infty} 2^{-\gamma L} = \frac{B}{1 - 2^{-\gamma}} = \frac{H_1 - h}{1 - 2^{-\gamma}} \tag{2.18}$$

Using E is more standard than γ. Moreover, E includes the "normalization" from h and H_1 so that we are told not just about the speed of the approach of h_L to h but also the distance from the start H_1 ($= H_1 - 0 = H_1 - H_0 = h_1$) to the end h.

In Fig. 12 we plot E as a function of h so that we can more easily grasp their joint behavior[32]. From

[32]This is called an entropy-complexity diagram [107, 27]

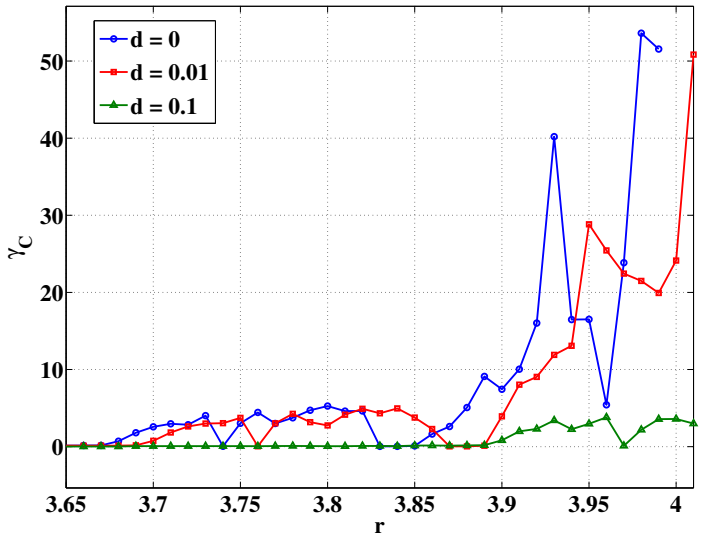


Figure 11: Decay rate of mutual information for logistic map and modified logistic map as a function of r using a binary partition ($d = 0$: ◯, $d = 0.01$: □, $d = 0.1$: △). The decay rate here was calculated as the reciprocal of the area underneath the mutual information curve: $1/\sum_0^\infty I(\Delta)$. If this is large, then the correlations are weak. As the coupling d increases, the correlations get stronger. Also, for $d = 0$ and $r = 4$, $\gamma_C \to \infty$.

Fig. 12 we also learn about the period doubling that occurs before the transition to chaos, since $E = \log_2 T$, where T is the period of the string. During the period-doubling the logistic map is periodic, so $h = 0$, but E is increasing[33]. By the time the transition to chaos has occurred, we are at the peak value of E. From then on E decays to zero as h goes to 1. That is, as the logistic map becomes increasingly unpredictable the correlations and thus the reduction in unpredictability also decreases. It is strange that the data series tells us the most about itself just after its transition to chaos.

We will always find this same behavior for any system, at long as h is close enough to 1. Of course, these behaviors are built into the definitions of h and E. Although we see this approach to $(h, E) = (1, 0)$ happening roughly monotonically with an increasing nonlinearity parameter r, we will find just the opposite with the nonlinearity parameter of turbulence Re (see Section 4.6).

So what have we learned from using these information tools with a dynamical system? By virtue of Pesin's Theorem we are assured to at least be able to say as much as other standard tools, but this isn't very satisfying. Through a combination of h and E we were able to characterize the logistic map's behavior

[33]Incidentally, this also points out that E is not a single-valued function of h. We must often look back to the functional dependence of E and h on r to fully understand what is going on. This is somewhat like parametrizing a circle with an angle, for which the two direction x and y are also not single-valued functions of each other.

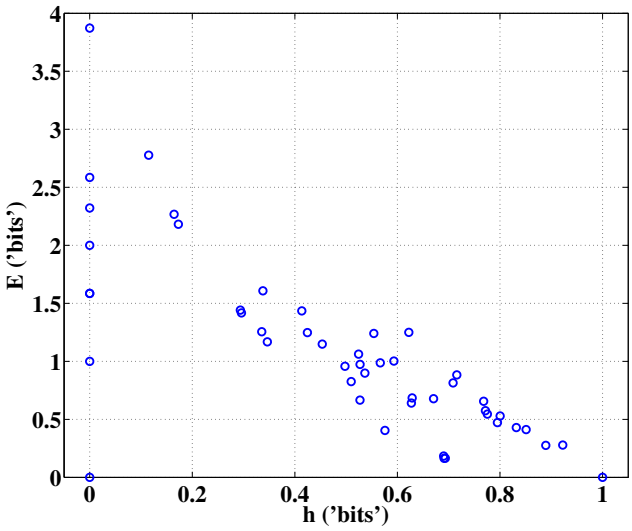


Figure 12: The excess entropy E plotted as a function of h. Now with E, we can understand more fully the behavior of the logistic map.

in a revealing way. There is more to the logistic map than just a route to chaos. Even in the chaotic regime the behavior of the logistic map evolves through increased randomness and decreased system memory until it is pure white noise. Using h and E we can appreciate a wider variety of chaotic behavior.

3.0 COMPUTATIONAL MECHANICS AND COMPLEXITY

We now introduce an offshoot of information theory and statistical mechanics: computational mechanics. The alleged purpose of this field is to describe a physical system's statistical properties in terms of how it stores and processes information [5]. Here we will only follow the school developed by James Crutchfield and his students [5, 107].

The main idea here is to create a statistical description, a model, of a system. Information theory is the language of description. That is, we use probabilities and entropies as the quantities characterizing the model's states and their evolution. This differs from a raw information theoretic treatment since that approach is limited to determining how much information is present, shared, etc. It is mostly descriptive. Computational mechanics goes a step further by providing a mechanism for this information state of affairs. This is something like the difference between kinematics and dynamics in mechanics.

3.1 STATISTICAL COMPLEXITY

Consider a stationary data set (or message) $X = (...X_{i-1}, X_i, X_{i+1}, ...)$, where $\mathcal{X}$ represents the set of the English syllabary or the range of velocity values in some experiment. The individual members of $\mathcal{X}$, denoted as usual by little x, have some probability distribution $p(X)$. Information from the past $\overleftarrow{X} = (...X_{i-3}, X_{i-2}, X_{i-1})$ is transferred to the future $\overrightarrow{X} = (X_i, X_{i+1}, X_{i+2}...)$ through the present. We will determine a set of special states called *causal states* S [5]. These will make up a minimal representation of our system for predictive purposes. Essentially, we are trying to build the simplest model possible.

To determine S, consider all unique blocks of data $X^{(L)}$, *i.e.* a sub-string of adjacent data values of length L. One would like to make L as large as possible, but the finite amount of data means only finite L can be statistically reliable. In practice, only a finite L is necessary anyway. With a large amount of data, we can calculate the conditional probability that any one past block $\overleftarrow{x^{(L)}}$ will give rise to any other future block of the same length: $p(\overrightarrow{X^{(L)}}|\overleftarrow{x^{(L)}})$. If the conditional probability distributions of any future blocks $\overleftarrow{X^{(L)}}$ are the same, the two past blocks are said to cause the same future. From a statistically predictive point of view the two past blocks are equivalent. Thus block 1 and block 2 are equivalent, $x_1^{(L)} \sim x_2^{(L)}$, if $p(\overrightarrow{X^{(L)}}|\overleftarrow{x^{(L)}}_1) = p(\overrightarrow{X^{(L)}}|\overleftarrow{x^{(L)}}_2)$.

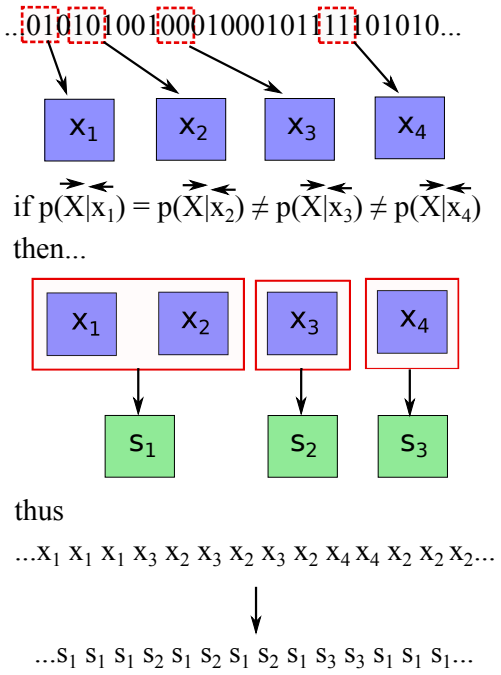


Figure 13: Causal states determined for $L = 2$ from a binary string of data with (initially) 4 distinct states. (The probabilities are not actually determined from the above string.)

All equivalent blocks are then combined into a single causal state. This is caricatured in Fig. 13. The states present in the data are combined (or not) after comparing their conditional probability distributions. We now have a set of predictive causal states S. The Shannon information (entropy) contained in S is the *statistical complexity*

$$C = H[S] = -\sum_s p(s) \log_2 p(s), \tag{3.1}$$

where the logarithm is taken to the base 2, giving units of "bits". C is the amount of information contained in our model. That is, it is the information needed to statistically predict the behavior of our system.

Here is how this will work in practice: we find the causal states S as just described and we also have determined the transition probabilities between the states S. Start out in some state x belonging to a particular s. Roll your dice to determine the next s using the known transition probabilities, and roll the dice again to determine a particular x belonging to this s. (We also know the probabilities of each x belonging to any s.) Expressed symbolically:

$$x \xrightarrow{x \in s} s \xrightarrow{p(s'|s)} s' \xrightarrow{p(x'|s')} x'.$$

Then repeat. In this way the data is reproduced in a statistical sense.

We needed to know an amount of information $C = H[S]$ to carry out the above prediction program. That is, we need to ask (on average) C "yes" or "no" questions in order to find the current state of the system,

and then predict from there. By design, this connects with the system's predictability, since organizing the message's parts into causal states will affect the value of C.

3.1.1 A simple example

As a simple illustration of these ideas, consider a coin flipping experiment where each subsequent flip (either a 1 or 0) will be the same as the previous one with probability $P \in [0, 1]$ [16]. This is just like a correlated random walk (CRW) [152].

Suppose first that $P = 0.5$. Then we have the usual fair coin toss experiment, with $h = 1$ and $C = E = 0$, since this system is maximally uncertain but statistically simple to predict with no information being shared between the past and future. We can work through the example to see that this is so.

As described above we must determine the conditional probabilities to determine the causal states. Let us do so for both $L = 1$ and $L = 2$. We will see that since this system is by construction (order-1) Markovian, we only need to consider $L = 1$. The following tables (Tables 1 and 2) show the conditional probabilities $p(\overrightarrow{x^{(L)}}|\overleftarrow{x^{(L)}})$ where the column is the future $\overrightarrow{x^{(L)}}$ and the row is the past $\overleftarrow{x^{(L)}}$. That is, the column is conditioned on the row.

$L = 1$	0	1
0	0.5	0.5
1	0.5	0.5

Table 1: The conditional probabilities (columns conditioned on rows) of the future on the past for a history of length $L = 1$. For a fair coin ($P = 0.5$), all of these values are the same.

$L = 2$	00	01	10	11
00	0.25	0.25	0.25	0.25
01	0.25	0.25	0.25	0.25
10	0.25	0.25	0.25	0.25
11	0.25	0.25	0.25	0.25

Table 2: The conditional probabilities (columns conditioned on rows) of the future on the past for a history of length $L = 2$. For a fair coin ($P = 0.5$), all of these values are the same.

Since for any of the pasts for either $L = 1$ or $L = 2$ we have respectively the same conditional probability distribution of the future, all of these states are not predictively different. So for both $L = 1$ and $L = 2$

separately we combine the pasts in either case into a single causal state. Of course, with only one causual state $p(s) = 1$ and so $C = 0$ automatically (see Eq. 3.1).

Now consider this same model when $P \neq 0.5$. If P deviates from 0.5 even slightly, $C = 1$ since we will always need to know 1 bit of information (the previous coin flip) to predict the future. We can show this for general P. Table 3 shows the resulting conditional probability distributions.

The probability of a 1 or 0 is 0.5 regardless of the value of P ($\neq$ 0 or 1). Eventually there will be a switch from 0 to 1 or 1 to 0, and for a very long string there will be just as many 0s as 1s. This means that the joint probability $p(00) = 0.5p = p(11)$ and $p(01) = 0.5(1-p) = p(10)$. The conditional probabilities are calculated as $p(0|0) = p(00)/p(0) = 0.5p/0.5 = p$, etc. The rest of the results are shown in Table 3. We can also calculate the block entropies H_L to get h and E (see Eq. 2.3 and 2.12). The results are $h = -P\log_2 P - (1-P)\log_2(1-P)$ and $E = 1 + P\log_2 P + (1-P)\log_2(1-P) = 1 - h$.

$L = 1$	0	1
0	P	1-P
1	1-P	P

Table 3: The conditional probabilities (columns conditioned on rows) of the future on the past for a history of length $L = 1$. Our coin's future is now coupled to its past by construction and so the conditional probability distributions are different than in the fair coin case.

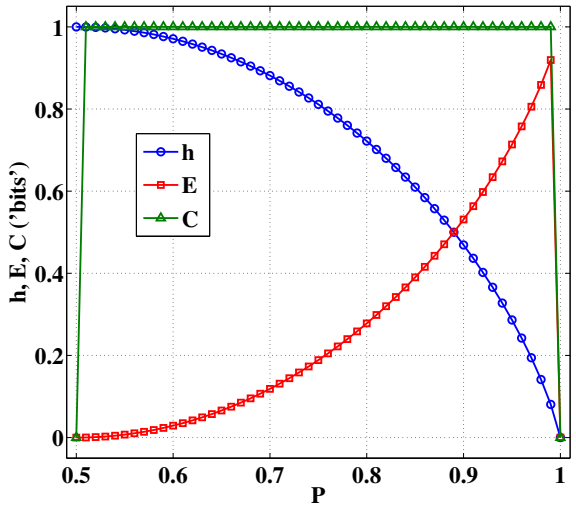


Figure 14: Plot of the fundamental quantities h, E and C for the simple example given here.

For a small deviation from 0.5 of $P = 0.55$, $h \simeq 0.95$ and $E \simeq 0.05$. Thus, if our system is "nearly" random with only weak correlations, h and E hardly deviate at all from their random values ($h = 1$, $E = 0$).

This example also highlights the difference between E and C [21, 20]. As we have mentioned previously, C is bounded below by E. It is a unique feature of this system that the difference is $C - E = h$. It is related to the system being first-order Markovian, *i.e.* each flip depends directly on the previous flip only and not any before that. We will avoid further discussion of this simple example until we discuss various interpretations about the relationships between C, E and h below.

3.2 INTERPRETATIONS OF C AND E

There are currently two main interpretations of C and E and their difference (if any). We also provide a third pragmatic interpretation.

3.2.1 Crutchfield

The first interpretation is of course due to Crutchfield and his students who developed this formalism. In their papers we can see an evolution of the interpretation of C which started out as only a complexity quantifier and evolved into the amount of information our system needs to compute [11, 5, 21, 22].

In their view, we have some process with which we associate a total amount of information C at any particular time. An amount E of this information is what the past and future have in common. Since $C \geq E$, we have another quantity $\chi = C - E$, which is hidden in the sense that E is the information we now know from looking at the infinite past. That is, if we see the whole message, we can know E, but we will not know χ unless we try to model the process with C and figure out its internal structure. For the simple example above, $\chi = h$ simply means that our uncertainty is always hidden from us. Keep in mind however, that it is not always true that $\chi = h$. It does seem to be true in general, however, that $\chi \geq h$.

The only problem with this interpretation is that it is difficult to interpret physically. We don't know who holds the information. Is it just a description or does the system itself truly process information? If we are thinking of a book, then it is not difficult to imagine the author as always having an amount of information C in mind as he/she writes, connecting what was written before any point with what comes after to an extent E and keeping the reader in the dark as to where the book will go by an amount χ. Moreover, the reader will be uncertain by an amount h at each page. We will need to take Landauer's interpretation (and that of Bennett, *etc.* as well, see Section 2.4.3), where nature computes, we are left with some unease about the interpretation of these quantities. This, of course, may be fine.

3.2.2 Wiesner

This recent and lucid interpretation is due to K. Wiesner et al. [37]. Just as with Crutchfield's interpretation, here we also must assume that information has a physical existence, as Landauer and others have (see Section

2.4.3). However, we need not go as far as they did in thinking about the universe as a computer. Instead, we consider the physical limitations on computation and modeling.

In this case, our computer needs an amount of information C to make its calculations. Calling E the predictive information, this is the amount of information stored in the computer that is transmitted through the computed data. This means that at each computational step, we carry on further an amount E and not C, which means we discard or erase $H_{erase} = C - E$ at each step. Wiesner et al. [37] find that any computational cycle must satisfy

$$H_{erase} + E - C = 0 \tag{3.2}$$

where H_{erase} is the information erased per step. Obviously, $H_{erase} = \chi$ from Crutchfield's interpretation. Here, however, we don't think of this as hidden but as unused, wasted information. Wiesner et al. [37] show that the thermodynamic cost of running the computation is kTH_{erase}. In the above example $H_{erase} = h$. In other words, since we throw away an amount of information h that we got in the previous step, we get back a "new" amount of information h on this next step!

They go on to define the predictive efficiency $E/C = 1 - H_{erase}/C$, in a way suggestive of a Carnot cycle. We put in an amount of information C and get out an amount E. This is the fraction of the information that actually affects the future.

While this interpretation is less vague than Crutchfield's, it is not clear whether it should be applied to physical systems either. Wiesner *et al.* do so with a version of Maxwell's demon [37], but as already mentioned this may not be meaningful (see Section 2.4.3).

3.2.3 Predictability

Our third interpretation is much more pragmatic and less adventuresome than the previous two. We ask: how predictable is a physical system? We start by assuming that information theory and computational mechanics provide us with useful tools for quantifying the uncertainty and certainty of a system's behavior. We have already shown how the uncertainty h and the Lyapunov exponents λ are related (see Section 2.4.1). Although λ (and by extension h) are often used to discuss the predictability of physical systems such as the weather, these only tell us about the *unpredictability* of these systems [5].

The difference between predictability and unpredictability can be understood with an example. Consider a completely random system such as an unbiased coin flip. We already know that here $h = 1$ is a maximum and $C = 0$ is a minimum. That is, the unpredictability as measured by h is maximum and the predictability is also a maximum. Here we think of predictability as being inversely proportional to C since this is the amount of information needed to predict. So we see that unpredictability and predictability are not just inversely related. In fact, as one can see from the plot of C vs. h for the logistic map in a later section (Fig. 19), C will generally behave non-monotonically with respect to h. This must be so, since $C = 0$ for $h = 0$ and 1.

If we do limit ourselves to a statistical treatment, we typically look for the "future" mean value of some variable with its standard deviation taken as a kind of error estimate [137]. However, if we know the probability of all events possible, we are able to calculate any moment (mean, variance, etc.). This means that an amount of information C is how much we need to know in order to calculate all statistical moments of our variable from any initial condition. This gives C a very useful interpretation in terms of predictability.

With regard to C's relationship with E and h, we already know that C is bounded below by E. In order to make an accurate prediction, we at the very least need to know an amount of information E, which is the information from all the correlations and structure in the data stream. At the same time, we cannot reduce our uncertainty below an amount h. Statistical prediction means we are not certain of our prediction, but we should know exactly how uncertain we are.

3.3 DIFFICULTIES IN ESTIMATING C

When actually handling real data to identify S, we must deal with imperfections. These imperfections may be due to external noise or the finiteness of the amount of data, *etc.* Regardless of the origin, some threshold needs to be set to determine if two conditional probability distributions for the future are the same, since they will never be identical.

An example of some conditional probability distributions is shown in Fig. 15. Two of the distributions are similar, indicating that the two states belong to the same causal state. The third distribution is entirely different. The task is to choose a sensible metric to make this distinction objectively.

This problem was apparently also recognized by the creators of this complexity machinery, although reference to it is apparently absent in previous literature [5, 11, 21, 20, 19, 22]. They developed the Causal State Splitting Reconstruction (CSSR: pronounced "scissor") algorithm, which is outlined in Ref. [23] and is available for download from its author's website [24].

We developed a MATLAB program that uses either a χ^2 or a Kolmogorov-Smirnov (KS) test to compare any two conditional probability distributions and determine S [25]. A 0.95 confidence level is used, but the results are not sensitive to this parameter. The CSSR algorithm and our own method are in good agreement, as can be seen from Fig. 16. We end up choosing the χ^2 method using our own code because it is simpler and the KS test in principle requires a continuous distribution. In the end, of course, choosing a suitable metric for determining the causal states has an element of subjectivity to it.

Another difficulty in estimating C comes in choosing a block size L. Here the problem of finite data we found earlier with h (see Section 2.3) comes back to haunt us. Observe Fig. 17, which is a caricature of what we would expect the behavior of $C(L)$ to be in the ideal case. Until L is large enough, we keep capturing more and more of the data's correlated structure. At some point $C(L)$ should stop increasing since beyond a certain L we will find no more interesting inter-symbol relationships. This should be called C, which is

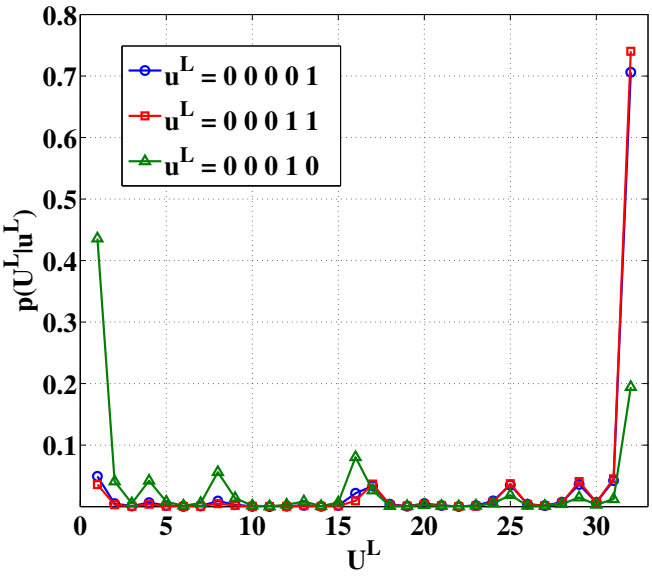


Figure 15: An example of three conditional pdfs used to determine the causal states. The data used here is binarized turbulence data with $L = 5$ (giving a total of 32 possible states) and $Re = 3300$ ($\bigcirc = 00001$, $+ = 00011$, $\triangle = 00010$). The horizontal axis features all the possible future states while the vertical axis is the conditional probability that given a certain past state, any of those possible future states will occur. Here the distribution for states $\bigcirc$ and $\square$ appear similar while that for state $\triangle$ is quite different.

independent of L. As we increase L further, we see the opposite behavior that we did to h after its leveling out. Since there is not enough data to accurately determine the probabilities, the data looks less random. Since C (generally) decreases as randomness increases, $C(L)$ increases with L.

Take a look at Fig. 18. In this figure we have plotted $C(L)$ vs. L for several different data streams. The periodic and random data sets behave as expected. Unfortunately, we do not see, or it is difficult to perceive, the flattening kind of behavior for the $C(L)$ determined for turbulent data as we did with h_L (see Section 2.3). It is difficult to know why, since the calculation of C involves many more steps than h does. We have come up with a simple criterion for choosing an L to determine C.

Since we always calculate h (and E) for the same data that we calculate C for, we know where the h_L flatten out and we have already chosen an L in that case. Roughly speaking, this means that the important features of the data are captured for an L of this length[1] and so we use the same L for C. We can also calculate h from the causal states

$$h = H(\overrightarrow{X}^1|S^{(L)}) \tag{3.3}$$

which we do as a simple test to see if our length L is sufficient. In reality, however, this many not differ at

[1]There could be a difference. See Ref. [20].

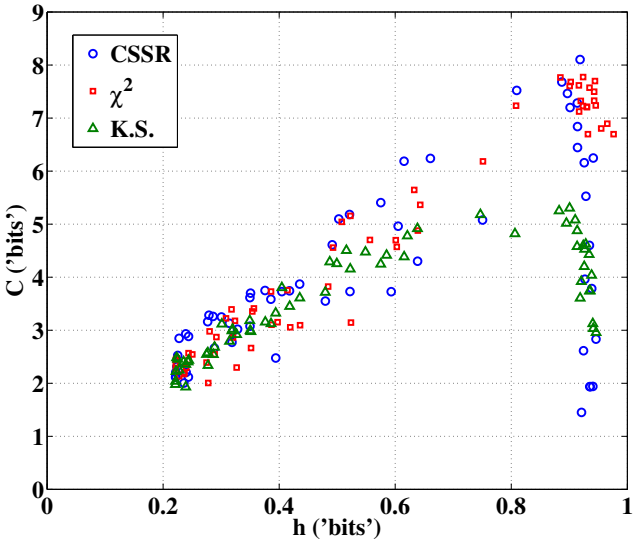


Figure 16: A comparison of several different methods of determining $C(10)$ for binarized turbulence data. The general trend is the same for all, but there are distinct differences.

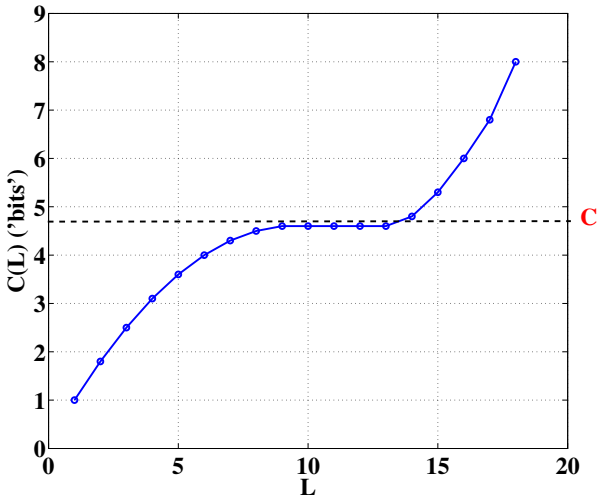


Figure 17: A schematic of the presumed behavior of C vs. L.

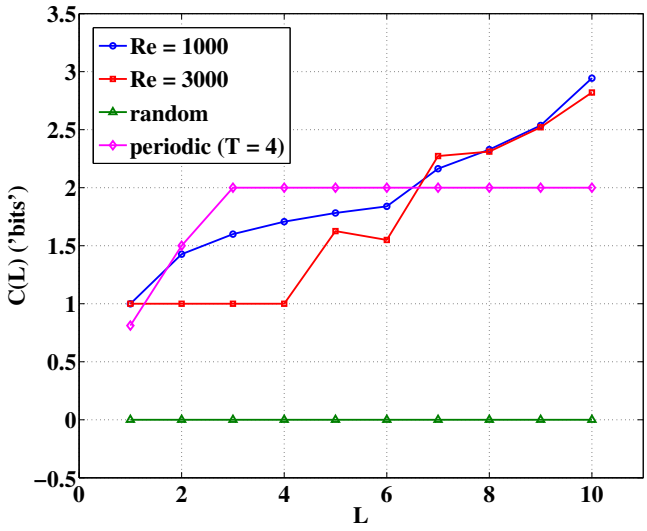


Figure 18: The real behavior of C vs. L for binarized turbulence data ($Re \simeq 1000$ (◯), $Re \simeq 3000$ (□)), random (△) and periodic data (◇). For the turbulence data there is no clear flattening out where we can determine the L-independent C. For the random data C is flat for all L, while for the periodic data it soon equals $\log_2(4) = 2$ bits. All data was analyzed using our MATLAB program for a dataset of size 10^6.

all from using the same L to determine C and h. Although we have not done so, we might also test some of the asymptotic properties of the causal states [12]. This may afford some additional quantitative tests.

3.4 COMPLEXITY OF THE LOGISTIC MAP

Just as we have analyzed the Logistic map using h and E, we will do so again including C as well. Now with these together we can potentially say even more, especially considering the possible interpretations of these tools discussed above.

Take a look at Fig. 19. We see that E and C behave slightly differently as functions of h. By definition they must both be 0 at $h = 0$, 1, and so it is natural for both to reach a maximum somewhere in between. Moreover, there is a gap between E and C between these values. We can try to "interpret" this result using the three ideas discussed above.

3.4.0.1 Crutchfield Interpretation: The chaotic behavior of the logistic map is cryptic or hidden. Until we build a model and uncover its structure (as revealed by Fig. 19), we won't understand how it works. It is not clear, however, why there should be hidden information for the logistic map or what this means. However, we may be able to become more comfortable with this if we draw an analogy. Consider

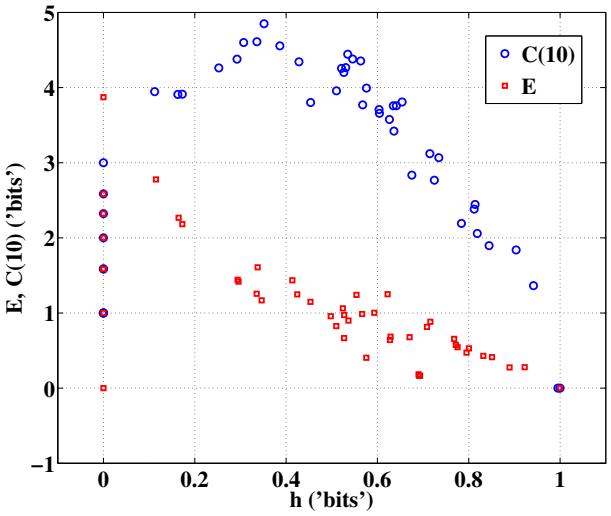


Figure 19: The quantities E and C plotted vs. h for the logistic map. This is a nice, parameter-free way of viewing the behavior of the system. As noted in the interpretation, $C > E$ is expected for $h \neq 0, 1$.

how our brain uses its memories. We cannot make decisions based only on our memories, we need to judge and interpet them and apply them to the situation at hand. This requires additional input (information) and may be like the χ in this situation.

3.4.0.2 Wiesner Interpretation: The thermodynamic cost of computing (simulating) a data series for the chaotic logistic map is *increasing* as $k_B T(C - E)$). If we were to run a simulation of the logistic map on a computer, this is the minimum amount of energy that will be dissipated at each step. Because $k_B T$ is so small, this theoretical limit is not likely to be of practical importance, but it is interesting.

3.4.0.3 Predictability Interpretation: If we were to make predictions of our system for intermediate values of h, the amount of information required to do so is actually higher than the amount of information we know about the future given the past. That is, having an infinite amount of experience of the system's behavior isn't enough to tell us what to predict next. We need to make a model.

Why do we need to make a model for prediction[2]? Simply put, E only tells us how much uncertainty we no longer have, but not how to make a prediction that achieves this. As far as this author knows, the only systems that have $E = C$ are (a) random, (b) static and (c) periodic. The first two are trivial, but (c) is interesting. The information containted in the fact that the data is periodic with a period T is $E = \log_2 T$. This same information is all that is needed to determine predict C with with no uncertainty whatsoever,

[2]Actually, we will always need to make a model to perform the prediction itself.

i.e., $h = 0$.

So what does that tell us about why we need a model? Just as with the simple example in Sec. 3.1.1, in general $C > E$ when $h > 0$ unless $h = 1$ (random). The general idea is again that prediction includes a mix of certainty and uncertainty about the future. This will be true for nearly all (interesting) systems. By how much C is greater than E, and how this difference depends on h will require additional information tools and/or theoretical considerations that depend on the particular system in question.

4.0 TURBULENCE

The word "turbulence" may remind us of a bumpy plane ride or a difficult time in our life. Its definition in fluid mechanics and physics is likewise uncomfortably uncertain [35, 40, 43]. Rather than worry too much about making a proper definition, researchers typically limit themselves to listing characteristics that "most" turbulent flows have [43]. For now, let us simply say that turbulence refers to a state of fluid flow in which the unsteady velocity field is fluctuating in both time and space.

We will use the Reynolds number Re as the main parameter characterizing the turbulent systems we discuss. This dimensionless parameter is determined by a length L and a velocity V that characterize the flow in some important way, along with the fluid's kinematic viscosity: $Re = VL/\nu$ [35, 40]. An example of this length and velocity might be the radius of a pipe $L = R$ and the mean velocity over its cross-section $V = \frac{1}{\pi R^2}\int_0^R \int_0^{2\pi} v(r,\phi) r dr d\phi$ [35, 40].

The Navier Stoke's equation is a continuum expression of the conservation of momentum for fluid "particles"

$$\frac{\partial \vec{u}}{\partial t} + (\vec{u}\cdot\vec{\nabla})\vec{u} = -\frac{1}{\rho}\vec{\nabla}p + \nu\nabla^2\vec{u} + \vec{f} \tag{4.1}$$

where $\vec{u} = \vec{u}(\vec{x},t)$ is the fluid velocity vector, $\rho = \rho(\vec{x},t)$ is the density, $p = p(\vec{x},t)$ is the pressure, ν is the kinematic viscosity and $\vec{f}$ is the body force (if any). Looking at this equation, we see that Re is the ratio of the inertial $(\vec{u}\cdot\vec{\nabla})\vec{u}$ and viscous terms $\nu\nabla^2\vec{u}$ [40][1]. Our interpretation of Re follows from this observation. At small values of Re, viscous damping forces dominate. At large values the fluid's own momentum (inertia) dominantly controls its behavior[2] with negligible influence from viscous damping.

We are mostly concerned with turbulence as it appears in two-dimensional (2D) systems, but we can learn some useful ideas from its conceptually simpler three-dimensional (3D) counterpart. For both 2D and 3D turbulence we will focus on the high Re, "fully-developed" case. We will also pretend for the most part that we are at a stage where the flow is Re-independent.

[1] This is if we make estimates of the terms based on the characteristic dimensional quantities V and L. Thus $(\vec{u}\cdot\vec{\nabla})\vec{u} \simeq \frac{1}{L}V^2$ and $\nu\nabla^2\vec{u} \simeq \nu\frac{1}{L^2}V$.

[2] This feedback mechanism is the the nonlinearity that makes turbulence such a difficult and interesting problem.

4.1 THREE-DIMENSIONAL TURBULENCE

Consider a large bucket of water. If we take our hand and stir the water we create motion at a scale roughly the size of our hand. We can check that we have done so by measuring the velocity as a function of position in the bucket and calculating the power spectrum of the velocity fluctuations.

$$E(k) = \delta u(k)\delta u^*(k)$$

where k is the wavenumber (inverse length) and $\delta u(k)$ is the Fourier transform of the velocity fluctuation $\delta u(x) = u(x) - \langle u(x)\rangle$. (I have ignored the vector nature of u and x for simplicity.) We call $E(k)$ the energy spectrum, since $\frac{1}{2}u'^2 = \int_0^\infty E(k)dk$ where u' is the rms velocity. That is, $E(k)dk$ is the energy per unit mass between wavenumbers k and $k + dk$. So if our hand is of size L, then $E(k)$ will have a peak around $k \simeq 1/L$ as in Fig. 20[3].

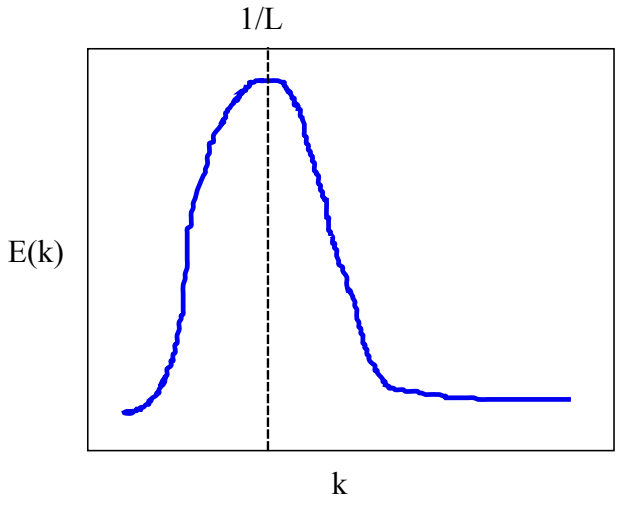


Figure 20: A cartoon of the energy spectrum for the water bucket stirred by our hand. Here the L is the size of our hand and the peak is at $k \simeq 1/L$.

We know from experience and experiments that the energy will not stay at $k \simeq 1/L$. If we take our hand out and watch, we will see the bigger motions change into smaller ones until they disappear and the water is still. This tendency for large scale motions or "eddies" to transfer down scale and finally dissipate into heat is one of the most prominent features of turbulence. Actually, it is perhaps the most important. It is the one characteristic nearly all discussions of turbulence eventually relate to. This is called the energy cascade.

Some claim that we can see from his paintings that Leonardo di Vinci saw and perhaps "understood" this. However, it was L. F. Richardson who first expressed this in words, borrowing from the poetry of Jonathan Swift [49]:

[3]One can do slightly better than this by contriving a cloud of vortices with a characteristic scale L and then looking at $E(k)$ [35].

> Big whorls have little whorls that feed on their velocity,
> and little whorls have lesser whorls and so on to viscosity.

This expresses quite nicely what we hopefully were able to imagine above from the bucket. Richardson worked on the problem of turbulent diffusion [50]. Besides developing the above intuition, he also was the first to realize that it is velocity differences separated by some distance in space that are the crucial quantities of interest.

What we are able to visualize and even express qualitatively with words, A. N. Kolmogorov codified in mathematical formulae. Just as with information we started with a qualitative idea and then came up with a definition based on a few simple axioms, starting from a few, simple assumptions Kolmogorov also showed that there is a downscale cascade of energy and that the energy spectrum must even take on a specific form[4].

We will follow closely the detailed presentation of Kolmogorov's arguments from the excellent book on turbulence by Frisch [51]. We first define a few quantities in order to proceed. The energy dissipation rate $\epsilon \equiv \langle 2\nu S_{ij} S_{ij} \rangle$, with $S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$ being the symmetric strain rate tensor, is the rate at which energy is transformed into heat. The large scale at which energy is injected into the flow is L, *e.g.* the size of our hand stirring the bucket.

Kolmogorov's first universality assumption is as follows [51]:

> At very high, but not infinite Reynolds number, all the small-scale statistical properties are uniquely and universally determined by the scale L, the mean energy dissipation rate ϵ and the viscosity ν.

We can then determine the form of the energy spectrum from dimensional arguments.

$$E(k) = \epsilon^{2/3} k^{-5/3} F\left[\left(\frac{\nu^3}{\epsilon}\right)^{1/4} k, Lk\right] = \epsilon^{2/3} k^{-5/3} F(\eta k, Lk)^5 \tag{4.2}$$

where we have assigned a new quantity often called the Kolmogorov or dissipative scale $\eta = \left(\frac{\nu^3}{\epsilon}\right)^{1/4}$. We will soon see that η is roughly where the cascade stops and dissipation begins. Since we are considering small scales, $Lk \to \infty$ and we assume that

$$F(\eta k, \infty) = F(\eta k) \tag{4.3}$$

Now for Kolmogorov's second assumption [51]:

> In the limit of inifite Reynolds number, all the small-scale statistical properties are uniquely and universally determined by the scale L and the mean energy dissipation rate ϵ.

In other words, he drops the dependence on ν since taking Re to infinity and ν to zero are equivalent. This assumption, along with the assumption of complete similarity [138] give us

$$F(0) = const. = C \tag{4.4}$$

[4] Unlike for H, Kolmogorov's assumptions are constantly debated since there are small but universal deviations from his result. However, the magnitude of the step he made can not be overstated.

[5] Frisch does not include the second argument, although there seems to be an additional step from Kolmogorov's assumption to do so [51].

and so

$$E(k) = C\epsilon^{2/3}k^{-5/3} \tag{4.5}$$

This is sometimes called the -5/3 law. Experiments have confirmed this scaling prediction repeatedly, to the point that the -5/3 spectrum is usually presented and thought of as proof that a fluid system is turbulent.

We have shown part of what we advertised, that Kolmogorov predicted the form of $E(k)$. But what about the direction and or existence of a cascade? Here we switch to the real space description of turbulence. We define the qth order longitudinal velocity difference on a scale r (also called the qth order longitudinal velocity structure function) as

$$S_q(r) = \delta u_q(r) = \langle(\vec{u}(\vec{x}+\vec{r}) - \vec{u}(\vec{x}))\cdot\hat{r})^q\rangle_x. \tag{4.6}$$

These are meant to follow their namesake and elucidate the structure of the turbulent flow. Unfortunately, their interpretation is a little muddled, but that is not important here (see Davidson [35]). For example, one can show that for isotropic turbulence, $S_2(r)$ and $E(k)$ are related (in 3D) by trigonometric transforms [35]. Also, this same quantity can be rewritten as $S_2(r) = 2[u'^2 - \langle u_{||}(x+r)u_{||}(x)\rangle_x] = 2[u'^2 - C(r)]$ where $C(r) = \langle u_{||}(x+r)u_{||}(x)\rangle_x$ is the spatial autocorrelation function of the longitudinal velocity $u_{||} = \vec{u}\cdot\hat{r}$. This makes an explicit connection between the correlation structure of the flow and $S_2(r)$.

Returning to the cascade, we will here deviate from Frisch's presentation since he was concerned with the issue of forcing. We can take the Navier-Stoke's equation (Eq. 4.1) for $\vec{u}(\vec{x})$ (call this NS$[\vec{u}(\vec{x})]$) and left multiply by $\vec{u}(\vec{x}+\vec{r})$ and add this to NS$[\vec{u}(\vec{x}+\vec{r})]$) to get the Karman-Howarth equation after some rearranging [35]. Assuming isotropy, we get a governing equation for the structure functions (and presumably the energy of the turbulent "eddies"):

$$-\frac{2}{3}\epsilon r^4 - \frac{r^4}{2}\frac{\partial}{\partial t}S_2(r) = \frac{\partial}{\partial r}\left[\frac{r^4}{6}S_3(r)\right] - \nu\frac{\partial}{\partial r}\left[r^4\frac{\partial}{\partial r}S_2(r)\right]. \tag{4.7}$$

Kolmogorov then took the infinite Re limit to ignore the ν term and also assumed a steady state so that we ignore the $\frac{\partial}{\partial t}$ term. This leaves

$$S_3(r) = -\frac{4}{5}\epsilon r \tag{4.8}$$

which is his famous "four-fifths law". Even without understanding what this means, this is a very attractive and simple result. It is also not difficult (in principle) to verify experimentally. Nevertheless, let us try to understand what it means.

First, from dimensional considerations, we see that $S_3(r)$ has units of $u^2 \times u$, *i.e.* of kinetic energy (per unit mass) flux. Because the velocities are actually velocity differences on a scale r, the flux should be thought of as coming into or out of a control volume of size r which can represent a typical "eddy" of that size. This flux is here related to the dissipation rate, as one would expect for the cascade. Let us now see about the minus sign above.

Consider a control volume of size r. The energy per unit mass in an eddy of this size is $\delta u(r)^2$. The change in energy (per unit mass) in the volume is given by

$$\frac{d}{dt}\int \delta u(r)^2 dV = \int \frac{d}{dt}\delta u(r)^2 dV = \int \epsilon dV \propto \epsilon r^3.$$

If we think of the usual continuum conservation laws, this change in energy must be due to some flux of energy in or out

$$\epsilon r^3 \simeq \int \delta u(r)^2 \delta u(r) dA \propto \pm \delta u(r)^3 r^2$$

or

$$\delta u(r)^3 \propto \pm \epsilon r$$

which gives us the same exponents as above. So what about the minus sign? The minus sign means that the flux is inward, *i.e.* the change is from energy coming from large scale into smaller scales. Thus the "four-fifths law" is a statement that energy cascades from large scales to small scales at a constant rate of ϵ.

All together, we have the following phenomenological picture. Energy is injected at a large scale L and at a rate ϵ. Energy is transferred from scale to scale and then finally dissipated at this same average rate. Using dimensional analysis and with the added evidence of the "four-fifths law" we expect that all the velocity fluctuations go as $\delta u_q(r) \propto \epsilon^{q/3} r^{q/3}$. So the "Reynolds number" associated with each scale is roughly $Re_r \simeq \delta u_1(r) r/\nu \simeq \epsilon^{1/3} r^{4/3}/\nu$. If we set $Re_r = 1$, we find that $r = \eta = (\nu^3/\epsilon)^{1/4}$, which is why we mentioned earlier that this scale is the "end" of the cascade: when Re is of order unity, viscous dissipation starts to dominate. Before this, the Reynolds number of each scale is large and we assume lossless transfer of energy downscale. One can also show that $L/\eta \propto Re_L^{3/4}$. In other words, the span of scales involved in the cascade increases with Re_L, the large scale Reynolds number[6]. This span of scales is called the "inertial range", since for these scales the Reynolds number is large, which means that the inertial term is dominating.

This phenomenological picture has had tremendous success. Even now, a large part of turbulence research deals with trying to understand the deviations from these ideas. This is the subject of intermittency.

4.2 TWO-DIMENSIONAL TURBULENCE

It was not too long after Kolmogorov's success in 1941 that theoreticians also became interested in 2D turbulence. Onsager made some contributions as early as 1949 [53]. However, the form of the theory that we are most used to now began with Kraichnan's work in 1967 [54] followed by Batchelor in 1969 [55], despite his earlier opinion that it was not interesting [56]. Theoretical and computational work was followed by experiments in the past two decades [33]. Now that we are familiar with the phenomenology of 3D turbulence, it will not be difficult now to proceed and understand 2D turbulence.

Turbulent flow in 2D is decidedly different from its 3D counterpart. In 3D, eddies of size r break up becoming smaller and smaller. The total vorticity amplitude is amplified by local velocity gradients, a well-studied effect called "vortex stretching" [51]. In 2D turbulence, on the other hand, the total squared vorticity $\langle \omega^2 \rangle$ (enstrophy) is a constant of the motion in the absence of viscosity, just like the total kinetic energy density $\langle u^2 \rangle$ in 3D. It was R. Kraichnan who first recognized that this new conservation law comes

[6] One can estimate this as $Re_L = u' L/\nu$.

into play with the result that energy is transferred to large scales while enstrophy is handed down to small scales [54, 59, 33, 32, 60].

Assume that the turbulence is forced at a scale $r = L$. In the steady state this forcing results in a mean rate of kinetic dissipation per unit mass $\epsilon = d\langle u^2\rangle/dt$ and a mean rate of dissipation of the enstrophy, $\beta = d\langle \omega^2\rangle/dt$. The Kraichnan analysis gives the following result for the second-order structure functions

$$S_2(r) \equiv \langle \delta u(r)^2 \rangle \propto \epsilon^{2/3} r^{2/3}, \quad r > L \tag{4.9a}$$

$$S_2(r) \equiv \langle \delta u(r)^2 \rangle \propto \beta^{2/3} r^2, \quad r < L. \tag{4.9b}$$

The two cascades do not extend to infinity in both directions. The energy cascade eventually reaches a scale on the order of the system size where it is cut off by boundary friction or coupling to the 3D environment [33]. Likewise, the enstrophy cascade eventually reaches a scale $r = \eta \equiv \nu^{1/2}/\beta^{1/6}$ where viscosity takes over.

In 3D, the sign of the third-order structure function is used as evidence of the cascade direction. However, there are potential problems with this interpretation. The reason is that there are actually contributions to $S_3(r)$ from enstrophy flux as well [35]. However, since one does not expect an enstrophy cascade in 3D, the analysis should be fine in this case.

In 2D, one expects that for an energy cascade inertial range, $S_3(r) > 0$, since energy goes to large scales. In fact, making the same assumptions as Kolmgorov for 2D, one is led to $S_3(r) = -\frac{3}{2}\epsilon r$. The enstrophy flux contributions to $S_3(r)$ are no longer insignificant (as they were in 3D) and we must look for another tool. In fact, the story is even more complicated [87] since the decay of 2D turbulence behaves differently than its 3D counterpart. (In 3D the effect of the decay of turbulence behind, *e.g.*, a grid does not affect $S_3(r)$ significantly).

All the complications aside for now, a simplified picture of the double cascade process is to view eddies of size $r > L$ progressively combining to create the energy cascade. The interstices between these eddies contain most of the shear (and hence large vorticity) but little energy. The result is a cascade of enstrophy to small scales. Recently this picture has come into question and an alternative mechanism has been proposed in Refs. [61, 62], but there is full agreement on the strong difference between 3D and 2D turbulence. Both of these cascades are studied in the experiments described here, although they do not appear simultaneously. This is the case in nearly all 2D experiments and simulations [33, 32, 60]. Some high resolution simulations and a few experiments are able to achieve the classical dual cascade picture [63, 64, 65, 66, 67].

Two dimensional turbulence occurs approximately in nearly all large scale weather patterns due to the fact that the thickness of the earth's atmosphere is very small compared to its breadth. How can we realize 2D turbulence in a laboratory experiment? The method used here is introduced in the next section.

4.3 EXPERIMENTAL SETUP

4.3.1 Soap Film Physics

In reality, there are no true 2D systems, turbulent or otherwise. All the systems which we treat as 2D are approximations that often have traces of the 3D world remaining. Usually, this approximation is very good. Our system is a soap film: a bubble that is drawn into a sheet. Bubbles can be made without soap, but the surfactants in soap endow the mixture with a self-healing property that makes these films very stable. A bubble of pure water is inherently unstable since surface energies would encourage it to collapse into a sphere. With a surfactant present, any thin region that begins to form tends to be filled. This is because the thin region will have a lower surfactant concentration, and the neighboring surfactant molecules will move in, dragging their water molecule neighbors. Moreoever, the surfactant makes the soap film behave like an elastic membrane, making it also stable with respect to out of plane motion.

The average thickness of the soap film is on the order of several μm. This can be measured by several methods, but perhaps the most precise has been an infrared transmission technique [32]. Water has a very large absorption peak at wavelengths near 3 μm. This means that when infrared light near this wavelength is sent through the soap film, it is strongly absorbed and the transmitted intensity can be used to determine the thickness of the soap film. It has been found that the thickness is very uniform for a turbulent soap film and fluctuations are less than 10% [32].

The thickness is many orders of magnitude smaller than the smallest turbulent scales present in experiments [33]. In this regard, this system is superior to the salt layer setups also used to study 2D turbluence [33]. However, just like these systems deal with the friction from the layers underneath, we must deal with air drag and Marangoni stresses. Air drag, the frictional losses due to the soap film pushing past the surrounding air, has been shown to play an important role in terms of energy dissipation [33, 94].

Marangoni stresses are caused by the surfactant. These stresses are due to essentially the same mechanism as the one responsible for stabilizing thin regions as described above. The soap molecules (surfactant) have an optimal spacing on the suface of the film, which translates into a specific concentration (number of soap molecules per unit area). All excess soap molecules will remain in the bulk of the film, perhaps collecting together to form micelles [153]. If a region of the soap film has a lower (higher) concentration than the optimal one, its neighboring molecules will sweep in (out) to restore the balance. This act of sweeping in drags fluid particles along which results in the Marangoni stresses. The effect of this on the fluid mechanics of the soap film is almost entirely unknown, except that it has been connected with a special kind of hydraulic jump in soap films [57].

The soap solution is a mixture of Dawn (2%) detergent soap and water. The exact recipe for Dawn soap is, of course, a trade secret. However, this presumably "dirty" soap (it is full of additives, *etc.*) was found to make soap films that were much more stable than purer surfactants. Small particles are also added for laser doppler velocimetry (LDV) measurements. Both the particles and the LDV measurements are discussed

next.

4.3.2 Velocity Measurements with the LDV

The LDV operates in just the way that its name suggests. A photon absorbed and re-emitted by a particle (discussed below) moving with the flow will have a frequency shift determined by its velocity relative to the source of the photon. With the speeds normally encountered in laboratory flows, the frequency shift is small and difficult to detect. For this reason, two coherent beams are carefully aligned so that the shift in frequency is not measured relative to the original frequency but can be measured directly. The simplest way to understand this is to imagine a particle crossing the two beams interference pattern like the one shown in Fig. 21.

A particle with travels through the interference pattern and the LDV will observe a large signal when the particle passes through a maximum and a small signal when the particle passes through a minimum. With a known spatial separation between maxima, the velocity of the particles is easily inferred from the signal's temporal separation. Of course this is only simple in principle. The technology developed for these measurements is complicated and incredible. For our experiments we used LDVs from both TSI, Inc. and MSE, Inc. For the purposes of the work done here there was no significant difference found between these two systems.

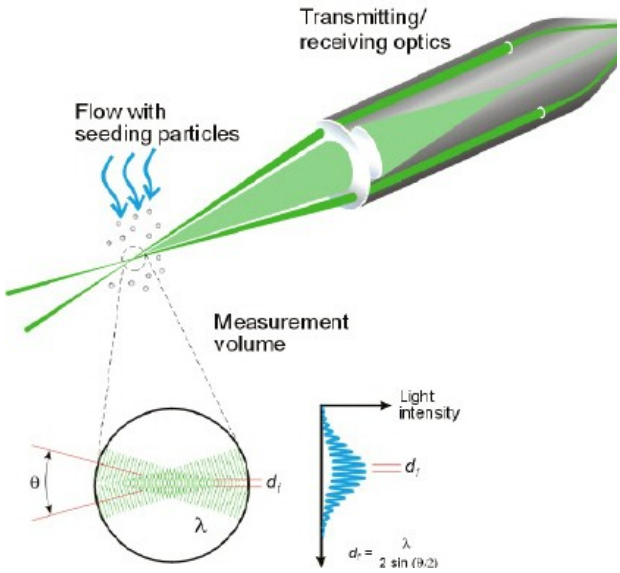


Figure 21: A schematic of the basic operation of a modern LDV [154].

There are several important parameters to keep in mind when choosing an LDV. First, there is some finite volume (probe volume) over which the measurements are made. The velocity measured is really an average over this volume. All LDV's have elongated, ellipsoidal probe volumes. While the diameter of this ellipsoid

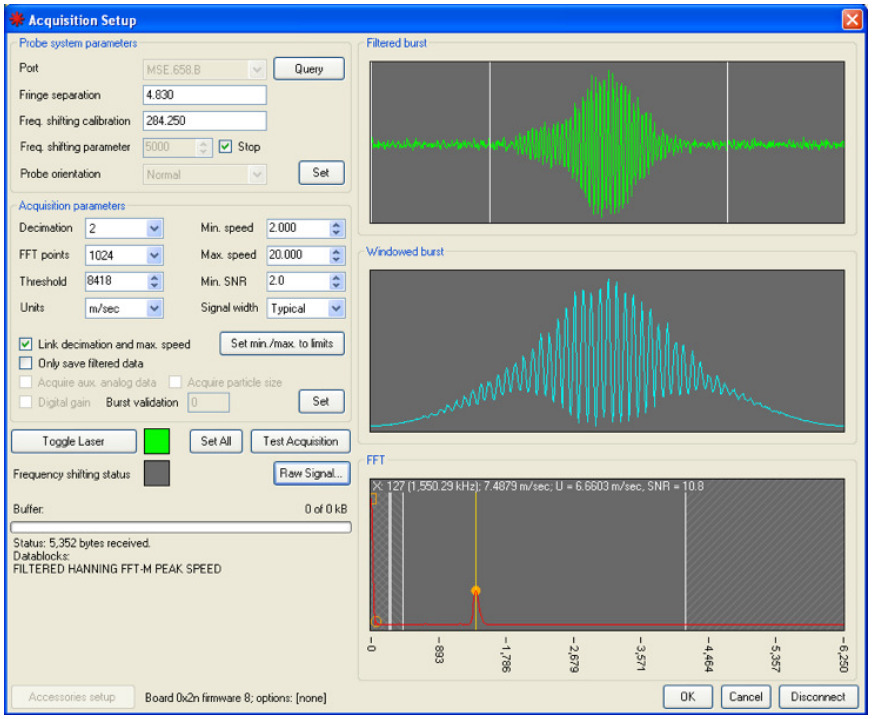


Figure 22: A screenshot from the MSE LDV software. The hardware recognizes a pulse (a burst of photons coming from a particle in the flow). The burst has an oscillation with a frequency corresponding to the speed of the particle. After filtering to reduce noise and windowing to focus attention only on the pulse, the frequency of the pulse is determined by taking the FFT (fast Fourier transform) of the signal. The power spectrum of the signal will be peaked at the frequency. The speed of the particle, which follows the flow, is inferred from this frequency.

may be on the order of 100 μm (smaller for TSI), its length is several mm. For the soap fim measurements, the smallest length scale we are typically interested in is only the order of several mm (based on estimates of the dissipative scale in *e.g.* the enstrophy cascade: $\eta \simeq \nu^{1/2}/\beta^{1/6}$.) Thus, the diameter of the probe volume is much smaller than the smallest length scale we are interested in, and since the soap film is roughly 10 μm thick, the several mm length of the probe volume is no problem at all. Moreover, the long length of the probe volume relative to the film thickness makes "focusing" the LDV at the desired measurement point all the easiser. (Incidentally, the thinness of the soap film also makes 2D PIV measurements simpler as well.)

Another feature to watch out for is the number of fringes in the probe volume. While it is best to have a small probe volume, this can also reduce the number of fringes. A large number of fringes means it is easier to resolve the frequency of the pulses used to determine the flow velocity. The number of fringes is determined by the wavelength of the laser light λ and the angle at which the two laser beams are set to cross

θ. The beam diameter is D and the fringe separation is

$$d = \frac{\lambda}{2\sin(\theta/2)} \tag{4.10}$$

which means that the number of fringes is

$$N \simeq D/d = \frac{2D\sin(\theta/2)}{\lambda} \tag{4.11}$$

Thus by increasing θ, one can increase N. However, this will mean that the focal distance from the measuring probe to the measurement volume is also reduced, since one is usually working with a fixed probe lens diameter. There can be simple experimental setup issues if the probe has to be placed too close to the measurement point, such as if the flow is encased in a chamber or is otherwise physically separate. Thus, one must balance many of these issues and determine what kind of LDV to choose based on what needs to be measured.

One final feature of an LDV that sets the more expensive models made by TSI and Dantec apart from others is how they create the frequency shifter. One of the beams in the LDV is typically shifted by a set frequency. With this done, the interference pattern shown in Fig. 21 will oscillate in the measurement volume. The reason this is done is so that the *sign* of the velocity may be determined, although it also allows for other filtering techniques around the shifted frequency. TSI and Dantec use a Bragg cell, which shifts the frequency with little noise added. A less expensive model such as that made by MSE shifts the frequency mechanically using a system of polarizers and waveplates (details are a trade secret). This introduces a contamination frequency into the signal. For the streamwise velocity measurements made here, we didn't need to turn on the frequency shifter and so our signal was not contaminated, but any measurements made with signals close to zero velocity were always suspect.

4.3.3 Seeding particles

In order for the LDV to work, there must be some optical inhomogeneity moving with the flow. Sometimes this can exist naturally, but in our case, we seed the soap film flow with glass particles. The particles are partially hollowed so that their effective density ρ_p is matched to that of water $\rho \simeq 1$ g/cm^3. The reason for this is that we need the particles to follow the flow as faithfully as possible [149]. If they are too dense, they will go their own way because of inertia and if they are not dense enough, they will experience buoyancy effects.

Another important parameter is the radius of the particle R. Obviously if this is too large, its movement will be influenced by the velocities averaged over its size. However, if it is too small, it will not scatter enough light to be used with the LDV. We need it to be much larger than the wavelength of light. Moreover, its response to the flow will be "sluggish" if it is too large. The response time of a particle is given by [149]

$$\tau_p = \frac{R^2}{3B\nu} \tag{4.12}$$

where B is a measure of the density difference between the particle and the fluid given by

$$B = \frac{3\rho}{\rho + 2\rho_p}. \tag{4.13}$$

Since the particles are density matched on purpose, $B \simeq 1$. The Stoke's number is then the ratio of this time scale to that of the smallest time scale in the flow, which in this case is the dissipative time scale τ_η, *i.e.* the Stokes's number is $St = \tau_p/\tau_\eta$ [149]. For the enstrophy cascade, this time scale is given by

$$\tau_\eta = \frac{1}{\beta^{1/3}}. \tag{4.14}$$

In our experiments we use particles of sizes 1.5 μm and 4 μm. Using the largest of these two, and estimating at its largest $\beta \simeq u'^3/w^3 \simeq (30 \text{ cm/s})^3/(1 \text{ cm})^3 \simeq 3 \times 10^4 s^{-3}$. Thus, at its maximum, we have $St \simeq (4\ \mu\text{m})^2 \times (30 \text{ cm/s}) \ / \ 0.01 \text{ cm}^2/\text{s} \simeq 5 \times 10^{-4} \ll 1$. So there is no chance of the inertia of the particles affecting the measurements.

Another issue may arise from the number density of the particles. If there are too many particles in the flow, they may on average be too close to each other and affect each other's motion. It is clear that the fluid will be affected by a particle on up to a distance on the order of the radius R of that particle. Suppose we put $M = 1$ g of 1.5 μm particles (a rather large amount), giving a volume of $V \simeq 1.4 \times 10^{-11}$, into a $V_0 = 1$ L $= 1000$ cm^3 container of soap solution. Then the average spacing of the particles is roughly $(V_0/(M/\rho_p V))^{1/3} \simeq 2 \times 10^{-3}$ cm. This is over 10 times larger than R, so there should be no effect from particles interacting with each other.

One last effect to consider is that of the electrostatic interaction of the particles. These particles are not charged, but the randomly induced dipoles in them can interact. However, the charges in the surrounding water molecules will serve to screen them, thus reducing the effective distance at which the particles may interact with each other. The typical means of estimating the distance at which charged particles can interact when surrounding by screening charges is given by the Debye length λ_D [150]. This will be an overestimate of the effective distance since these glass spheres are not charged but can have an induced dipole. In water at room temperature $\lambda_D \simeq 1\ \mu\text{m} = 10^{-4}$ cm [151]. Given the estimate above that the average spacing of the particles is roughly 10^{-3} cm, we are again safe from worrying about another damaging effect.

4.3.4 Details of setup and protocol

Figure 23 is a diagram of the experimental setup. The soap film is suspended between two vertical blades connected to a nozzle above and a weight below by nylon fishing wire. The nozzle is connected by tubes to a valve and a reservoir which is constantly replenished by a pump that brings the spent soap solution back up to the reservoir. The flow is gravity-driven. Typical centerline speeds $\overline{u}$ are several hundred cm/s with rms fluctuations u' ranging from roughly 1 to 30 cm/s. The channel width w is usually several cm. Not shown in the diagram is the overflow tube: to keep the pressure head (due to gravity) constant a hole is cut into the top reservoir and a tube empties into the bottom reservoir. The pump pumps fluid up at a higher rate

than goes into the soap film, but because of this overflow hole the pressure head is kept constant and the entire system is stationary.

The stability of the soap film is always an important issue. Strong air movement created by swinging doors or air vents will not only create out of plane motion in the soap film but will also cause it to break. For this reason, the soap film is sometimes placed in a tent or a protected space.

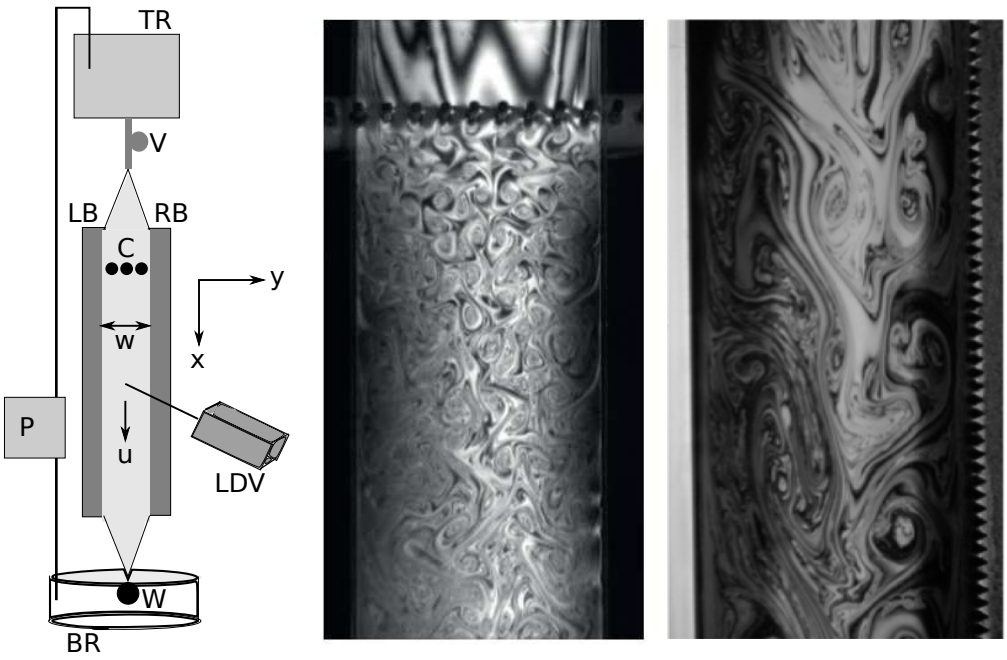


Figure 23: Left: Experimental setup showing the reservoirs (TR, BR), pump (P), valve (V), comb (C), blades (LB, RB), LDV and weight (W). Middle: Fluctuations in film thickness from turbulent velocity fluctuations with smooth walls and a comb. Right: Thickness fluctuations with smooth and rough wall. (Figure adapted from [31].)

We have several protocols for generating turbulence in the soap film. All of these amount to changing the boundary conditions on the flow, but the kind of 2D turbulence that is generated can be rather different.

We can either (1) insert a row of rods (comb) perpendicular to the film, (2) replace one or both smooth walls with rough walls (saw blades) with the comb removed and possibly a rod inserted, or (3) use a comb with smooth walls as in (1) but now very near the top of the soap film where the flow is still quite slow. When protocol (1) is used we almost always observe the direct enstrophy cascade [32, 33]. If procedure (2) is used, we can observe an inverse energy cascade [32, 33, 34]. When protocol (3) is used, we see no cascade at all.

The type of cascade is determined by calculating the one-dimensional velocity energy spectrum $\mathcal{E}(k)$,

where $\frac{1}{2}u'^2 = \int_0^\infty \mathcal{E}(k)dk$ and k is the wave number[7]. For the enstrophy cascade, $\mathcal{E}(k) \propto k^{-3}$ and for the energy cascade $\mathcal{E}(k) \propto k^{-5/3}$. In the protocol (3) case $\mathcal{E}(k)$ is flat and there is no power-law scaling and so apparently no cascade, although the flow is not truly laminar ($u' \neq 0$). These "no-cascade" measurements are used with the information theory analysis but not for the intermittency study. See Fig. 24 for some representative spectra.

The enstrophy cascade is extremely robust, in the sense that it is always observed with protocol (1) when near the comb. (Some distance down from the comb a scaling of $\mathcal{E}(k) \propto k^{-5/3}$ can be observed.) The ability to see the inverse energy cascade depends sensitively on the flux and channel width. Sometimes the enstrophy cascade is seen even when protocol (2) is used. We never saw a cascade when protocol (1) was used, however.

It is the unfortunate but unavoidable truth that 2D soap film turbulence is very sensitive to boundary conditions. The reasons are not at present understood. This presents an interesting avenue of research for the future.

As described above, one should in principle be able to also use $S_3(r)$ to classify the type of cascade. However, as also mentioned above, there are other complicating factors that make this difficult. Besides the effects of decay, perhaps the most damaging of these is that $S_3(r)$ does not only include contributions from energy being transported between scales but also enstrophy [35]. (This may not be such a large problem in 3D, but it is a potentially enormous problem in 2D where there is an enstrophy cascade.) It would be better to use the flux functions such as in Ref. [122]. However, this requires one to use measurements from particle imaging velocimetry (PIV) which were not available when this LDV data was acquired, so we will have to satisfy ourselves with using the spectrum $E(k)$.

4.3.5 Taylor's frozen turbulence hypothesis

Unless otherwise noted, we measure the longitudinal (streamwise) velocity component at the horizontal center of the channel. The data rate is $\simeq$ 5000 Hz and the time series typically had more than 10^6 data points. For this system, the time series is really a spatial series by virtue of Taylor's frozen turbulence hypothesis [35, 32, 33]. This is how we are able to calculate $\mathcal{E}(k)$ and $S_3(r)$, *etc.* which require a spatial series.

Taylor's frozen turbulence hypothesis is based on an observation concerning time scales. In a flow with a very large mean velocity, any turbulent fluctuation will be swept along (convected) very quickly. With an LDV, which measures the velocity at a single point, we obtain a time series of the velocity

$$u(t) = [...u(t - 2\Delta t), u(t - \Delta t), u(t), u(t + \Delta t), u(t + 2\Delta t), ...].$$

Refer to Fig. 25. The fluid particles with labels A and B are travelling at a mean speed U relative to

[7] How an (inverse) length is derived from a times series is discussed next.

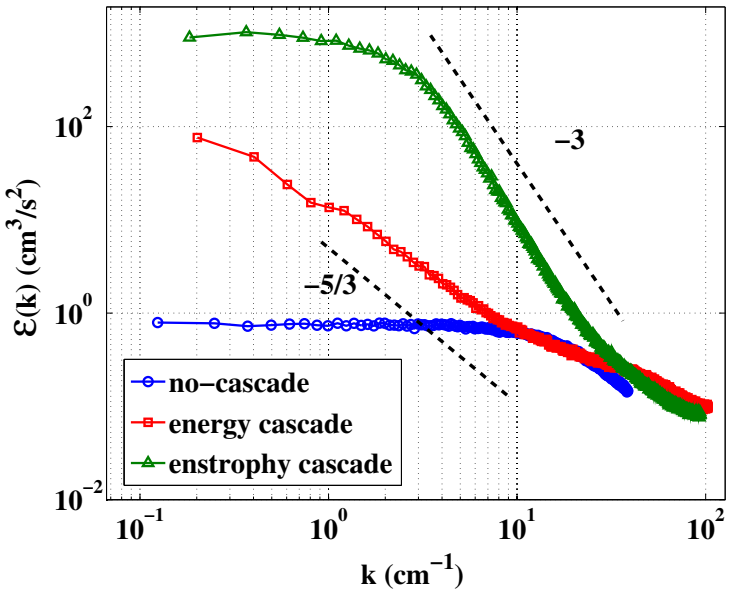


Figure 24: Representative one-dimensional energy spectra in a log-log plot of $E(k)$ vs. k. The enstrophy cascade (△) has a slope close to -3 while the energy cascade (□) has a slope close to -5/3. The flat curve (◯) is interpreted as belonging to data with no cascade.

the observation point. The velocity of the fluid particle B is measured at time t and a time Δt[8] later, the velocity of A is measured. The points A and B are clearly separated in space by a distance Δx. If their relative velocities are small, then $\Delta x = U\Delta t$. Another way of thinking about this is that the time scale of the eddies being carried along is long compared with Δt, so they have not changed in any way as they are shifted along the measurement point.

We can make an estimate to see if this is a reasonable assumption. If the rms velocity u' (a typical deviation from the mean U) is comparable to U, then particle A will have moved significantly more or less than $U\Delta t$ by the time it is measured. That is, if the turbulence intensity is $u'/U < 1$, we can use Taylor's frozen turbulence hypothesis. The applicability of this assumption have been tested in 2D soap films and works well up to a surprising $u'/U \sim 0.15$ [84]. Now the time series from above becomes

$$u(x) = [...u(x - 2\Delta x), u(x - \Delta x), u(x), u(x + \Delta x), u(x + 2\Delta x), ...].$$

Since the spatial and temporal behavior of turbulence is actually quite different, this is important to remember [58].

We emphasize that we measure a time series $u(t)$ which we must convert to a spatial series $u(x)$. Thus,

[8] Δt is is the inverse of the data rate, which is determined by many factors such as processor speed, particle density, light intensity, *etc.*

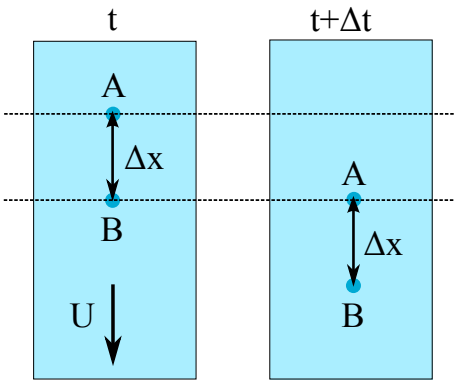


Figure 25: Caricature of a flow field in which a strong mean velocity U enables (forces) us to use Taylor's frozen turbulence hypothesis. The fluid particle A arrives at the measurement point a time Δt later than B. If their relative velocities are small enough, then this difference in time is simply translated into a difference in space $\Delta x = U\Delta t$.

out of the full temporal and spatial dependence of the velocity field $u(x,t)$, we only know about the x behavior.

4.3.6 Interpolation

The data are interpolated for equal spacing. To ensure insensitivity to the interpolation method, several are used. Linear and nearest neighbor schemes (similar to sample and hold [85]) are used in conjunction with the usual Taylor hypothesis: $x = U \times t$. The third method introduced by Kahalerras *et al.* seeks to correct for biases in the sampling stemming from the fluctuations in velocity [82]. No significant difference is found between these different interpolation schemes.

4.4 2D INTERMITTENCY

Now let us discuss intermittency, another name for the deviations from the description of 2D turbulence given previously. Of course, the reason this name is used is because, at least for 3D, the deviations are linked with strong, rare events or "intermittent bursts" in the velocity.

Intermittency is not a sharply defined effect in 2D or 3D. The presence of strong, unlikely velocity fluctuations is manifest in many different types of analysis. Most measures of intermittency are defined in terms of a deviation from the predictions of the scaling theory of Kolmogorov (K41) in 3D [68] or Kraichnan (Kr67) in 2D [54], although intermittency has a larger meaning than this [51]. Some of those measurements are made in the "inertial range" of velocity fluctuations, defined as those eddy sizes where damping effects

are negligible. The K41 and Kr67 predictions are for this range of scales. On the other hand, intermittency is reported in measurements of the probability density function (pdf) of velocity gradients $\partial_x u$ or velocity differences $\delta u(r)$. The pdfs are expected to be nearly gaussian if there is no intermittency, with large fluctuations being manifested in the tails. For values of r in the inertial range, the assumptions of K41 and Kr67 also require the shape of the pdfs to be independent of r, i.e. self-similar. The statistics of the velocity gradients provide a measure of the "smoothness" of the velocity field at very small scales outside the inertial range. For 3D energy and 2D enstrophy cascades, these scales lie in the dissipative (viscous) range.

Whereas intermittency is well-established for 3D turbulent flows, there is little agreement about its presence or origin (if present) in 2D turbulence. It is important to consider the two cascades separately. Numerical simulations by Boffetta *et al.* and experiments in salt layers by Paret and Tabeling suggest that the inverse energy cascade is intermittency-free [71, 72]. An experiment by Jun and Wu (JW) involving a horizontal soap film at a large Reynolds number indicates the opposite [73]. The soap film experiments of Daniel and Rutgers, which exhibit a dual cascade, indicate the presence of intermittency in both [66].

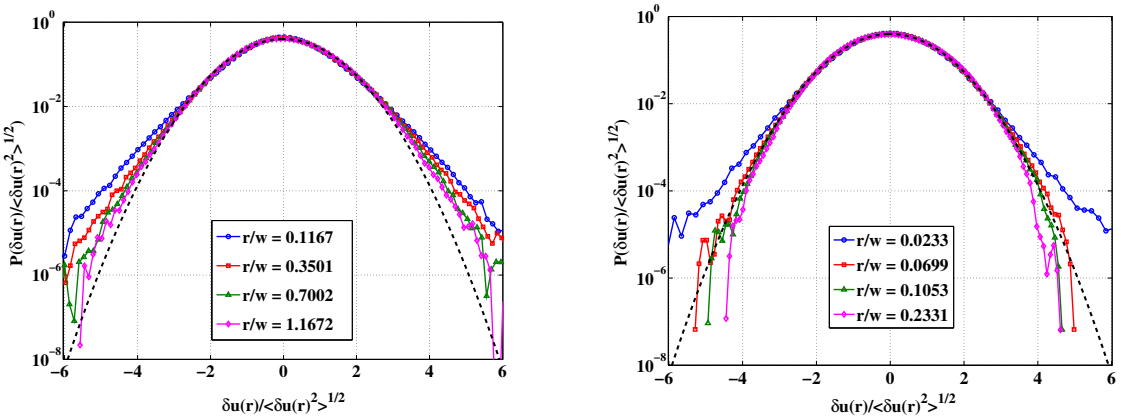
The role of 3D effects (in this case air drag) is also not clear. In some simulations it is held responsible for intermittent effects in the inertial scales of the enstrophy cascade [74, 75, 76]. The same linear drag force produces no such effect in numerical simulations of the inverse cascade and is necessary to establish a steady state [71, 33]. Moreover, there is negligible intermittency in salt layer experiments for either the energy or enstrophy cascade despite significant 3D effects from the floor of the container [72, 77]. However, a systematic study of the effects of drag on the enstrophy cascade indicate the opposite [78]. Other researchers claim that the dissipative scales of the enstrophy cascade are intermittency-free but that coherent structures emerge to produce inertial range intermittency [80, 81]. The possible effect of Marangoni stresses, which are present in soap films, is also unknown [57].

Here, the Reynolds number dependency of intermittency for 2D turbulence is discussed. The Reynolds number used here is the Taylor microscale Reynolds number R_λ, which ranges from 20 to 800. Strong intermittency is found for both the inverse energy and direct enstrophy cascades as measured by (a) the pdf of velocity differences $P(\delta u(r))$ at inertial scales r, (b) the kurtosis of $P(\partial_x u)$, and (c) the scaling of the so-called intermittency exponent μ, which is zero if intermittency is absent. Measures (b) and (c) are quantitative, while (a) is qualitative. Here we will only discuss (a) and (c). More details on all of these measures can be found in Ref. [58].

These measurements are in disagreement with some previous results but not all. The velocity derivatives are nongaussian at all R_λ but show signs of becoming gaussian as R_λ increases beyond the largest values that could be reached. The kurtosis of $P(\delta u(r))$ at various r indicates that the intermittency is scale dependent. The structure function scaling exponents also deviate strongly from the Kraichnan prediction. For the enstrophy cascade, the intermittency decreases as a power law in R_λ.

4.4.1 PDF of velocity differences

Figure 26 shows the normalized longitudinal velocity difference probability $P(\delta u(r)/\sqrt{\langle \delta u(r)^2 \rangle})$ as a function of the dimensionless velocity difference $\delta u(r)/\sqrt{\langle \delta u(r)^2 \rangle}$ for several values of r for both cascades. Both pdfs must be skewed and hence not perfectly gaussian to assure that the energy or enstrophy flux is nonzero [87]. It is apparent from this qualitative measure of intermittency that the energy cascade pdf has wider flanks than the enstrophy cascade pdf and is notably nongaussian. On the other hand, the enstrophy cascade pdf is well-fitted by a gaussian function (dashed line). The relative difference between the R_λ-values is significant. Some enstrophy data with smaller R_λ do not have a perfect gaussian distribution.



(a) Energy cascade with $R_\lambda = 50$.

(b) Enstrophy cascade with $R_\lambda = 610$.

Figure 26: The pdfs of $\delta u(r)$ measured for (a) the energy cascade at $R_\lambda = 50$ and (b) the enstrophy cascade at $R_\lambda = 610$. The first three values of r here are in the inertial range as determined by the power law scaling region of the structure functions. The dashed line is a gaussian function with zero mean and a standard deviation of unity. The mean and variance of the velocity data have been normalized so that if they are gaussian, they will lie on top of this curve. The pdfs at different r do not have the same shape, indicating a lack of self-similarity.

An interesting finding not discussed at length in Ref. [58] is that the pdfs of $\delta u(r)$ at large r for the enstrophy cascade data are sub-gaussian, thus, self-similarity is not restored as $r \to \infty$, just as in 3D turbulence [146]. This is the opposite trend of the small scales. Whereas large but rare fluctuations will fatten a pdf, a sub-gaussian pdf indicates some kind of constraint on the fluctuations. This has not been further investigated.

The pdfs at different r have measurably different shapes. This indicates a lack of self-similarity, since the statistics of velocity fluctuations should be the same at all scales r in the inertial range. The absence of self-similarity provides additional evidence that the flows are intermittent. The classical Kr67 theory of 2D turbulence (as well as the K41 theory of 3D turbulence) assumes self-similarity in the inertial range [54, 51].

Although these measures of intermittency are qualitative, they signal a disagreement with the most recent and prevalent view of 2D intermittency [33]. That is, it is held that the energy cascade should be non-intermittent while the enstrophy cascade is necessarily intermittent (for a variety of reasons). A more quantitative measures will be investigated next.

4.4.2 Intermittency exponents

One of the many measures of intermittency involves the structure functions $S_n(r)$. In the absence of intermittency, $S_n(r) \sim r^{\zeta_n}$, with $\zeta_n = n/3$ for the energy cascade (in 3D or 2D), according to K41 and Kr67 [68, 51, 54]. For the 2D enstrophy cascade, $\zeta_n = n$ [59]. Subsequent 3D experiments have shown that ζ_n/n is not a constant but decreases with increasing n. This is one identifier of intermittency [51]. The departure of the ζ_n from K41 and Kr67 is often measured using the "intermittency exponent" $\mu \equiv 2\zeta_3 - \zeta_6$, with $\mu = 0$ for perfect agreement with the theory.

Just as with the velocity derivatives and indeed any measure of intermittency, there are sometimes delicate and controversial issues involved in estimations. The issue of estimating scaling exponents will be discussed below, but before this can be done the reliability of the higher order structure functions should be addressed. As in the experiment of JW [73], the method advanced by Dudok de Wit is used to indicate the highest reliable order [91]. This method has the advantage of being less subjective than other methods, although the other checks give the same result. See the the appendices in Ref. [58] for a more complete discussion. Using Dudok de Wit's method, most data were reliable up to order 12 or 13.

The soap film experiments described here add evidence that intermittency is also present in 2D. The strongest prior evidence for its presence in 2D comes from a study of forced steady-state turbulence at R_λ comparable to that used here. In these experiments of JW [73], the 2D system is an electrically conducting soap film driven by an array of small magnets placed below it.

They evaluated the moments $S_n(r)$ using extended self-similarity (ESS) and several other methods as a check [92]. With ESS, one plots $S_n(r)$ as a function of $S_3(r)$ rather than r itself. With this approach the range of self-similarity extends over a broad range of the independent variable $S_3(r)$, as compared with a plot of $S_n(r)$ vs. r itself. The same has been done in this study. Figure 27 shows an example with $S_2(r)$, $S_4(r)$, and $S_6(r)$ vs. $S_3(r)$ for enstrophy cascade data with $R_\lambda = 490$. The dashed lines in the figure are least square fits to the data.

Figure 28 provides a comparison of the various methods for estimating ζ_n. One sees here the normalized structure function exponents ζ_n/ζ_3 for values of n in the interval 1 through 10. The open squares ($\square$) designate direct measurements of ζ_n/ζ_3 from the enstrophy cascade data ($R_\lambda = 490$) without invoking ESS. This method is labeled "Simple". The open circles ($\bigcirc$) are obtained using ESS, as discussed above. The open triangles ($\triangle$) are calculated as $\langle \frac{d \log S_n(r)}{d \log r} / \frac{d \log S_3(r)}{d \log r} \rangle$ just as in JW [73]. The straight dashed line is the Kr67 prediction and has a slope of $\zeta_n/\zeta_3 = n/3$ (no intermittency). It is apparent that the various methods give similar results without any significant differences occurring for larger n. The main results for

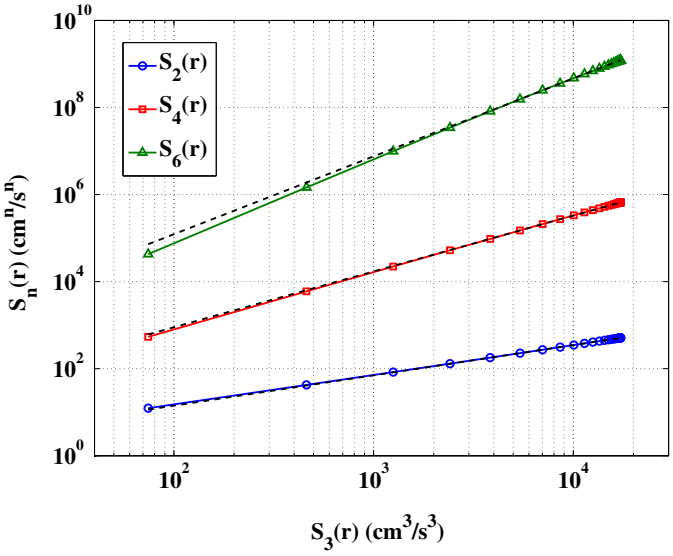


Figure 27: Three measured structure functions whose scaling exponents are determined using extended self-similarity (ESS). That is, $S_n(r)$ is plotted vs. $S_3(r)$. From the bottom to the top, the lines are $S_2(r)$, $S_4(r)$ and $S_6(r)$. Here $R_\lambda = 490$. These data belong to the enstrophy cascade.

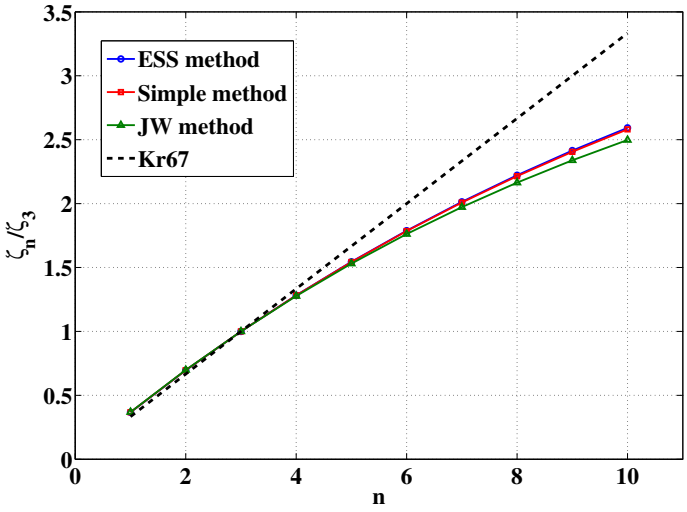


Figure 28: Normalized scaling exponents of the nth-order structure functions out to $n = 10$ for the enstrophy cascade with $R_\lambda = 490$. The data set denoted by squares (□) is extracted from measurements spanning a decade in r. The open circles (◯) denote slopes deduced using extended self-similarity [92]. The triangles (△) denote measurements obtained using the method from JW [73]. All methods agree very well with each other.

μ presented here were obtained using the ESS method.

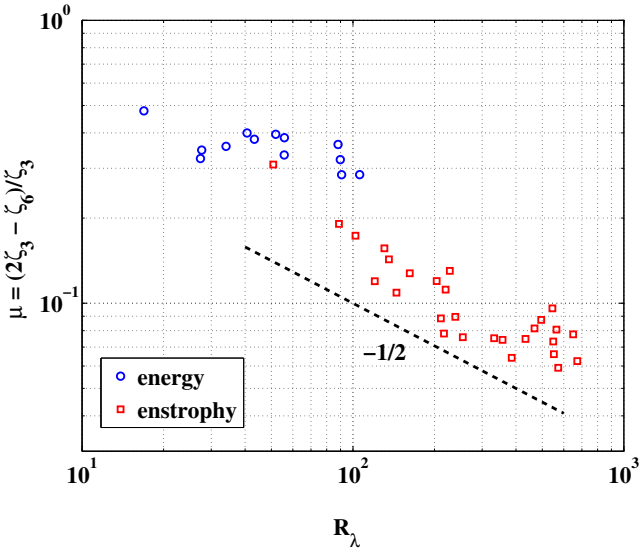


Figure 29: Intermittency exponent μ vs. R_λ for energy cascade (◯) and the enstrophy cascade (◯). The value of μ for the energy cascade is roughly constant while μ for the enstrophy cascade appears to be a decreasing function of R_λ.

In the present study μ is measured as a function of R_λ, which was not an adjustable parameter in prior measurements (JW's experiment varied R_λ but reported findings for only one value [73]). The results in Fig. 29 show that the dependence of μ on R_λ cannot be ignored. The observations are divided into two sets. The open circles (◯) denote measurements made for the energy cascade, where $10 < R_\lambda < 110$. The open squares (□) denote measurements made for the enstrophy cascade, where $50 < R_\lambda < 700$[9]. For the energy cascade, the maximum value of μ is $\simeq 0.5$, which is more than twice its value in 3D flows [93, 146]. It is notable that μ depends rather strongly on R_λ, at least for the enstrophy cascade. Just as with F_η, this decrease of μ with R_λ suggests that as $R_\lambda \to \infty$, intermittency vanishes for 2D turbulence.

4.4.3 The new picture of 2D intermittency

The values of the intermittency exponent μ measured here are significantly larger than that obtained in a salt layer for the energy cascade [72]. For this magnetically driven salt layer experiment, (also using ESS) $\mu \simeq 0.02$. The value of μ found by JW was $\mu \simeq 0.11$, which is comparable with the enstrophy cascade values measured here, although JW focused on the energy cascade [73]. The JW experiment quotes $R_\lambda \simeq 170$, which is significantly higher than the maximum value obtained in the present energy cascade experiments. This may explain the discrepancy in the measured values of μ. The soap film experiments which obtained

[9]The clear separation in R_λ between the two cascades is no accident, but it is also not fully understood. As was mentioned earlier, the energy cascade's appearance is sensitive to many conditions such as flux, distance from the comb, etc.

a dual cascade do not quote μ, but there is a stronger deviation in the exponents in the enstrophy cascade range than the energy cascade [66]. Their work has the novelty of looking at a dual cascade, which may be the reason for the disagreement with the present measurements. It may not be meaningful to compare the intermittency between experiments when R_λ is different. For most previous studies, no R_λ (or any Reynolds number) is quoted [71, 74, 76, 72, 77].

The argument that air drag is responsible for the intermittency in the enstrophy cascade does not deal with its R_λ-dependence. A term linearly proportional to the velocity (or vorticity) is usually added to the 2D Navier-Stokes equation to account for 3D drag effects:

$$\frac{\partial \mathbf{u}}{\partial t} + (\nabla \cdot \mathbf{u})\mathbf{u} = -\frac{1}{\rho}\nabla p + \nu \nabla^2 \mathbf{u} - \alpha \mathbf{u} + \mathbf{f} \tag{4.15}$$

where $\mathbf{f}$ is some external forcing and α is the drag coefficient [75, 76, 71, 74, 94]. The value of α extracted from a horizontal soap film experiment is $\alpha \simeq 0.7 \pm 0.3$ Hz [94], while estimates based on a boundary layer approximation suggest $\alpha \simeq 0.1$ Hz [33]. The corrections to the enstrophy cascade scaling exponents are on the order of $\alpha' = \alpha/\omega'$, where ω' is the rms vorticity [78, 74, 76, 75]. Namely, the exponents are modified to become $\zeta_n \simeq n(1-\alpha')$. This correction is estimated to be on the order of 10^{-3} for the system used here. This value of α' is in good agreement with that calculated in [33] for falling soap films. It is clearly too small to be detectable. Experiments with soap films in partial vacuums, where the air drag has been considerably reduced, indicate that the air drag has little effect on the spectral exponent [95]. The effects of air drag on the inverse energy cascade have not previously been considered since such 3D effects are necessary to maintain the steady state. With regard to coherent structures, their presence is always accompanied by energy spectra $E(k)$ that are much steeper than is observed in this work ($\gamma \geq 5$) [80], so this seems unlikely to be the root of the intermittency observed here.

In accord with some prior studies and in disagreement with others, the observed intermittency is as strong in 2D as in 3D. The measurements, which span more than two decades in R_λ, explore both the energy and enstrophy cascades. The intermittency is strongest in the energy cascade and is nearly R_λ-independent. While there is intermittency in the enstrophy cascade, it decreases with increasing R_λ, suggesting that it may vanish in the limit $R_\lambda \to \infty$.

The origin of the intermittency is not clear. There have been several explanations proposed in the literature for the direct enstrophy cascade, but they do not account for its strength or deal with its R_λ-dependence. There also seems to be no fundamental reason to expect the inverse energy cascade to be intermittency-free. Some of the difficulties in coming up with a unified picture of intermittency in 2D may also be due to the failure of the Kr67 picture. The present results suggest the need for taking a fresh look at the statistics of 2D turbulence, as some are already doing [61, 62, 89, 90].

4.5 INTERMITTENCY FROM AN INFORMATION-THEORETIC POINT OF VIEW

We can look at the above results in a slightly different way using the tools of information theory. Suppose for a moment that a turbulent cascade really exists in all the ways we described above. Thinking again of the bucket stirred by our hand, before we stir it is obvious that $H[p(\delta u(r))] = 0$ at any scale r. (For now let us consider the simplest quantity we have, H.) When we begin to stir, the information will necessarily increase from 0 at the scale of our hand $H[p(\delta u(L))] > 0$. Whether it is energy or enstrophy, the cascade takes this quantity which we injected at one scale and transfers it to another scale. This is not unlike a communication channel. Not only energy (or enstrophy) but perhaps information will also be transferred.

A perfect cascade has a constant transfer rate of its "stuff" (energy, enstrophy) from the injection scale to its dissipation scale[10]. Now consider the information. If the information present at injection is transmitted without error down to the dissipative scales, what kind of cascade would this be? Alternatively, suppose some information is garbled. What kind of cascade would this be? We will argue that a Kolmogorov cascade requires the information to be garbled, and that any that is communicated is responsible for the intermittency.

4.5.1 GOY Model

We first introduce and apply these ideas to a toy model which is designed to imitate a turbulent cascade. This makes it a nice testing ground for some of these ideas. The Gledzer-Ohkitani-Yamada (GOY) shell model is a set of complex coupled ODE's [139, 140]. Each variable (shell) corresponds to velocity fluctuations on a different spatial scale. These shells, however, live in Fourier space, and so their independent variable is the wavenumber k. (This is also why they are complex.) This shell model can be thought of as a truncation of the Navier-Stokes equation in Fourier space.

There are a finite N number of shells and we denote by n any particular shell (from 1 to N). We used $N = 22$ here. The wavenumbers k_n are picked to be a fixed logarithmic distance apart: $k_n = k_0 q^n$, where here $k_0 = 2^{-4}$ and $q = 2$. Large n corresponds to small scales, while small n refers to large scales. The governing set of equations is

$$\Big(\frac{d}{dt} + \nu k\Big) u_n(t) = i(a_n u_{n+1}(t) u_{n+2}(t) + b_n u_{n-1}(t) u_{n+1}(t) + c_n u_{n-1}(t) u_{n-2}(t))^* + f_n \tag{4.16}$$

where f_n is the forcing and a, b and c are shell dependent but constant in time. These constants determine the strength of energy flow between scales. They are

$$a_n = k_n \tag{4.17}$$

$$b_n = -\frac{1}{2} k_{n-1} \tag{4.18}$$

[10] This depends on how we interpret or model intermittency [51].

$$c_n = -\frac{1}{2}k_{n-2} \tag{4.19}$$

and following Ref. [139] we choose the forcing to be

$$f_n = \delta_{4,n}(1+i) \times 5 \times 10^{-3} \tag{4.20}$$

which acts only at the $n = 4$ shell. We followed the numerical scheme outlined in Ref. [139] and chose as our "viscosity" the values $\nu = 10^{-4}$, 10^{-5},10^{-6}, 10^{-7}. This can be thought of as spanning three decades in $Re \propto 1/\nu$. We ran the equations using MATLAB for roughly 200 turnover times (largest time scale) [139] and then computed all of the statistics.

The energy spectrum is given by

$$E(k_n) = \frac{\langle |u_n|^2 \rangle}{k_n} = \frac{e(k_n)}{k_n} \tag{4.21}$$

where $e(k_n)$ is the energy at scale k_n and the flux through each shell by

$$\Pi(k_n) = -Im\Big\langle \frac{k_n}{4}(u_{n-1}u_n u_{n+1} + 4u_n u_{n+1}u_{n+2})\Big\rangle \tag{4.22}$$

where Im denotes the imaginary part of the expression. Below we plot $E(k)$ in Fig. 30 and Fig. 31. Both

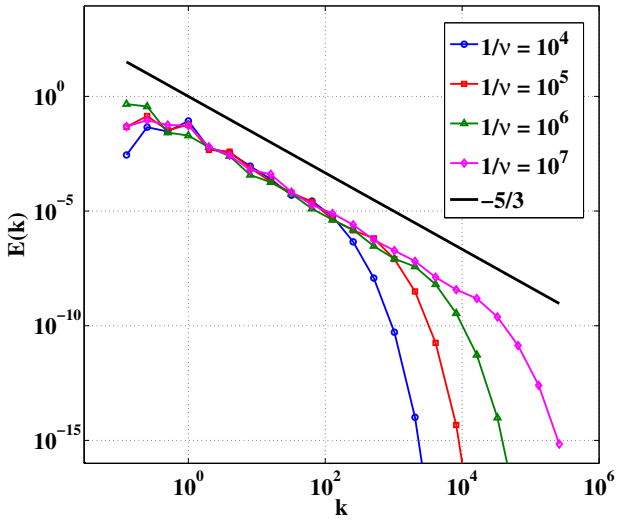


Figure 30: The energy spectra for various values of $1/\nu$ for the GOY model. The scaling of all curves is close to $E(k) \propto k^{-5/3}$ by construction. There is also a well-defined dissipative range for each, where $E(k)$ drops due to viscous effects.

these curves show behavior that idealized 3D turbulence shows: -5/3 scaling in $E(k)$ (with a nice dissipative range), and a constant region of $\Pi(k)$ which corresponds to the constant energy transfer rate between scales

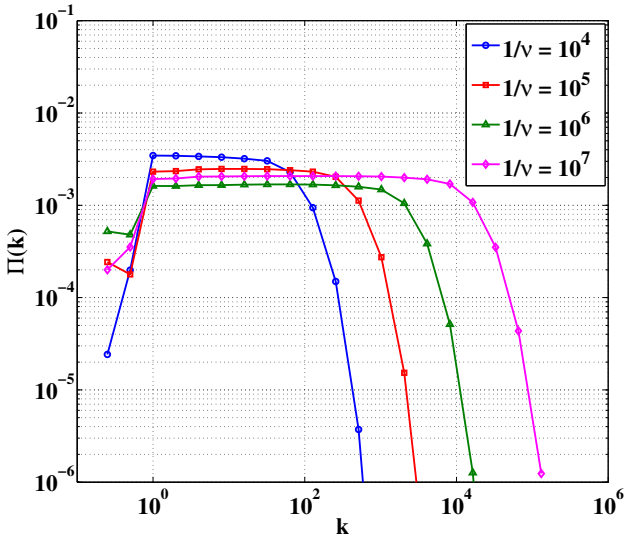


Figure 31: The energy flux through a shell k for various values of $1/\nu$ for the GOY model. A flat region in each curve corresponds to $\Pi = \epsilon$, the injection, transfer and dissipation rate.

(shells). As the viscosity increases, the dissipative range comes sooner in k. The inertial range, where $E(k) \propto k^{-5/3}$ and $\Pi = \epsilon =$ constant, is shortened.

This model also shows intermittency, as revealed by the "structure functions"

$$S_p(k_n) = \langle |u_n|^p \rangle. \tag{4.23}$$

One can show that this GOY shell model is energy-conserving in the inertial range (see Ref. [139]) and so the same dimensional arguments apply here as in 3D turbulence. The structure functions should scale as $S_p(k_n) \propto k_n^{-\zeta_p}$ with the exponents going as $\zeta_p = p/3$. This is not the case, as revealed by Fig. 32 shown below. We find that $\mu = 2\zeta_3 - \zeta_6 \sim 0.3$.

The information tools that we will use in this case are the mutual information of the energy between scales $I(e(k_n); e(k_n + \Delta k_n) = I(k; k + \Delta k)$ and the entropy $H(e(k_n)) = H(k)$. The mutual information is calculated for the same time t for each adjacent scale. Recall now the universality assumptions of Kolmogorov. From dimensional arguments we have

$$E(k) = \epsilon^{2/3} k^{-5/3} F\Big[\Big(\frac{\nu^3}{\epsilon}\Big)^{1/4} k, Lk\Big] = \epsilon^{2/3} k^{-5/3} F(\eta k, Lk)$$

and if we assume complete similarity [138], $Lk \to \infty$ gives us

$$F(\eta k, \infty) = F(\eta k).$$

The meaning of complete similarity here is that the small scales "forget" the large scales. That is, there is no (or little) information passed down from the large scales to the small. We can test this idea using

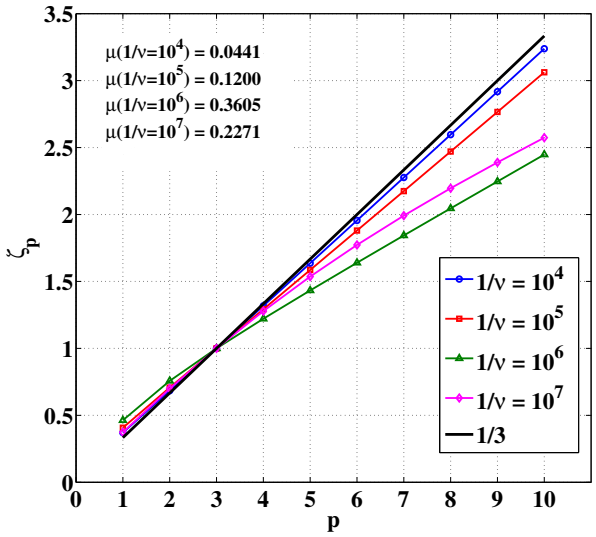


Figure 32: The scaling exponents ζ_p for the GOY model. There are clear departures from the expectations from dimensional arguments. Moreover, this departure seems to grow (non-monotonically) with $1/\nu$. The typical value of $\mu = 2\zeta_3 - \zeta_6$ is ~ 0.3.

information theory, in particular the mutual information. If one scale forgets about the other, than the mutual information between them will be small. Moreover, if they do not forget about each other, then we can not assume complete self-similarity. We might instead assume incomplete self-similarity, *i.e.* that $F(Lk, \eta k)$ retains its dependence on L in the form of

$$F(Lk, \eta k) = (Lk)^{\alpha} \widetilde{F}(\eta k) \tag{4.24}$$

where we have assumed that Lk is very large. The parameter α cannot be determined from dimensional analysis and must usually be determined experimentally or from extra assumptions [138]. For the turbulent structure functions, this translates to

$$S_p(r) \propto \epsilon^{p/3} r^{p/3} \left(\frac{r}{L}\right)^{z(p)} \tag{4.25}$$

where $z(p)$ is a (potentially p-dependent) deviation from the K41 result. We see that remembering and intermittency (exponents) are closely related.

We now discuss a premature idea to connect these ideas quantitatively. In the β model of intermittency [51], the cascade is imagined as a binary process of large eddies breaking up into smaller ones. The space filled by each successive step of the cascade is only a fraction of the previous step: $\beta = 2^{D-3}$, where D is the dimension of the phase space that turbulence actually resides in. As has already been mentioned, we can show that $\mu = 3 - D$, so it is related to the fraction of space that is not filled by turbulence. Instead of interpreting β as a fraction of space filled by turbulence, let us instead think of it as the ratio of forgotten

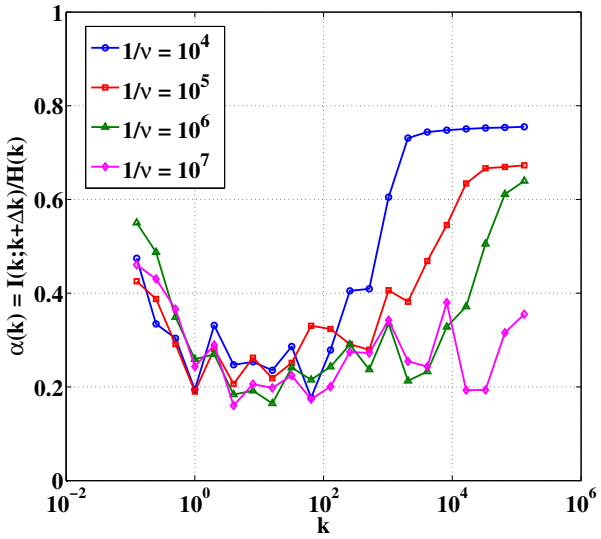


Figure 33: The ratio $\alpha(k) = \frac{I(k;k+\Delta k)}{H(k)}$ vs. k for several different values of $1/\nu$. This fraction of remembered information is smallest in the inertial range, as one might hope for us to successfully use complete similarity. The rough mean value for all cases in this range of k is $\alpha \sim 0.2$.

information to total information. Since the mutual information is the unforgotten information, and the entropy is the total information, we define

$$\alpha(k) = \frac{I(k; k + \Delta k)}{H(k)}. \tag{4.26}$$

This line of thinking thus suggests that $\beta = 1 - \alpha$. We plot $\alpha(k)$ for several values of $1/\nu$ in Fig. 33.

There isn't a very discernible trend in $1/\nu$, except that the value of $\alpha(k)$ seems to reach a progressively smaller value at large k (small scales) as $1/\nu$ increases. It seems that for large and small k the information is not forgotten. Let us take the value in the inertial range to be $\alpha(k) = \alpha \sim 0.2$[11]. In this case, we find that $\mu = 3 - D = -\log_2 \frac{1}{1-\alpha} \sim .322$, which is not far off the value that we calculated (see Fig. 32). However, we should not take this too seriously. The important point is that $\alpha(k)$ seems to reflect the existence of the cascade which nearly forgets.

4.5.2 2D Enstrophy Cascade

Let us now consider real data from our soap film. First we plot $S_2(r)$ for some enstrophy cascade data to see which scales are involved in the enstrophy cascade. From Fig. 34 it appears there is relatively good scaling for $10^{-2} < r/w < 10^{-1}$, where w is the channel width. The exponent is not always 2 as it should be from dimensional arguments, but this is a typical result [32, 33].

[11] This value should be independent of the partition size used since it is a ratio of two information quantities.

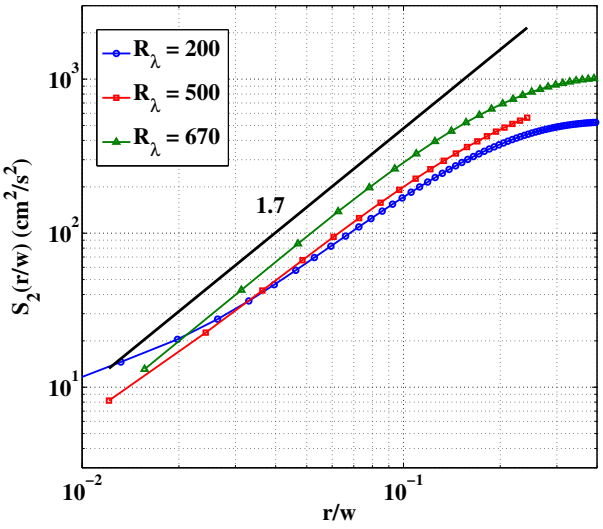


Figure 34: We plot the second-order structure function $S_2(r/w)$ for enstrophy cascade data to get an idea for the range of scales involved in the cascade. The power law scaling appears to be roughly in the range $10^{-2} < r/w < 10^{-1}$.

Just as with the GOY model, we calculate the ratio of the transferred or "remembered" information to the total information at each scale: $\alpha(r) = I(r; r + \Delta r)/H(r)$. This is plotted for several values of R_λ below in Fig. 35. Like the GOY model, the remembered information is less in the inertial range compared to larger

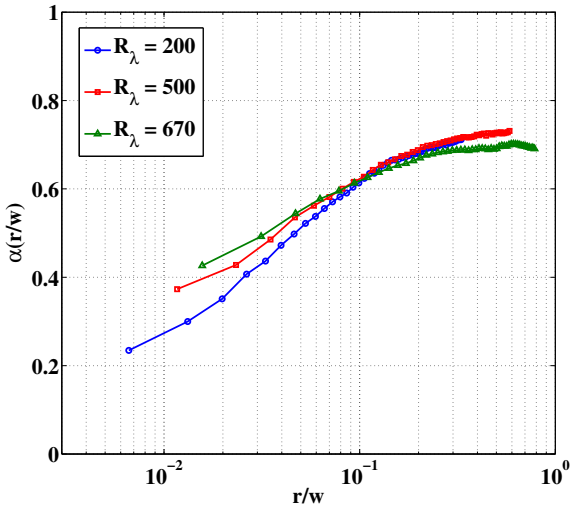


Figure 35: The ratio of remembered information to total information $\alpha(r) = I(r; r + \Delta r)/H(r)$. In the inertial range, $\alpha(r)$ decreases considerably from its large r value. However, as R_λ increases, the amount of remembered information increases instead of decreasing.

scales. However, here $\alpha(r)$ seems to be strongly r dependent, making an estimate of μ difficult. Moreover,

the value of α is larger for higher R_λ, which is the opposite of what we would have expected given that μ is a decreasing function of R_λ (see Fig. 29[12]. Although we have argued that an intermittency-free cascade must forget, it must also certainly be true that it is not a cascade if it remembers nothing. After all, the concept of a cascade inherently involves correlations between scales. It may be that a true cascade cannot be intermittency free. This issue of a cascade's forgetfulness has been addressed in a different way using the Fokker-Planck equation in Ref. [124].

Below we plot one other information theoretic quantity that can be used to quantify intermittency. This is the Kullback-Leibler divergence, where the reference probability distribution is a gaussian distribution with the same mean and standard deviation as the data. Fig. 36 shows that in the inertial range, the PDFs of the velocity differences are nearly gaussian, but they deviate strongly at the edges of the cascade and outside. In summary, the tools of information theory have rarely if ever been applied to the turbulent cascade. However, there are certain questions that seem perfectly suited to it.

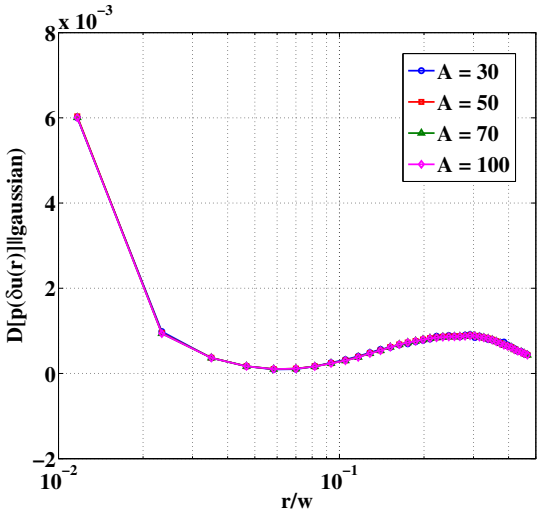


Figure 36: The Kullback-Leibler divergence between the probability distributions of the velocity differences and a gaussian distribution. A small value of D corresponds to the distribution being nearly gaussian, which is usually associated with a lack of intermittency. Several different alphabet sizes A were used here, although the result is independent of this choice. Here $R_\lambda = 500$.

4.6 ENTROPY OF TURBULENCE DATA

Let us now use the information tools that we have developed with 2D turbulence data. We will first do the simplest thing possible and calculate h for a stream of velocity data. Then we will discuss the interpretation of the results.

[12]Moreover, unlike with the GOY model, we are not able to calculate the mutual information for scales at the same time. This may be a confounding obstacle.

As far as we know, there have been no studies focusing on the spatial disorder or unpredictability of turbulence. Although turbulence is both a temporal and spatial phenomenon, the fundamental work of Kolmogorov deals only with the spatial structure of turbulence [68, 35]. The work of Kraichnan and others has shown that many of the essential features of turbulence are retained if one throws away temporal correlations but keeps spatial ones [98, 99, 100].

The temporal entropy rate h has been used in several novel experiments to show that the onset of turbulence in many systems can be described by a low-dimensional strange attractor [42, 102, 103, 104]. The evidence for this is provided by probing h at the transition. In the case of Taylor-Couette flow [102], h and the largest Lyapunov exponent λ increase with the Reynolds number Re. The usual expectation is that this trend continues as Re increases, as suggested by some models and analytic work [103, 104]. This may be true for h, but this work provides evidence that h follows a different path.

All the work on h or λ that we are aware of focused on either models of 2D and 3D turbulence [103] or only on the transition to turbulence [101, 102]. To our knowledge, no numerical simulations of the Navier-Stoke's equation have addressed this issue. Several studies of temporal predictability (actually unpredictability) had rough analytic relationships but then tested them with sophisticated models [129, 130]. Here we go well into the regime where turbulence has developed a cascade and exhibits the behavior we normally think of as "fully-developed". This may be the first experimental study of these ideas.

Presumably, the disorder h of fluid flow is relatively small if the flow is almost laminar. In this limit of small Reynolds number Re, one expects h to increase with Re. In the opposite limit of large Re, a so-called inertial range of correlated eddies of various sizes develops. Increased correlations implies added constraints or redundancies, which should lead to a decrease in h (see Sec. 2.1.5 and Ref. [4]). One therefore expects that after passing through a maximum, h will decrease with increasing Re. (This issue is addressed in Ref. [97] from a thermodynamic point of view.)

The present experiments probe a turbulent system (quasi-2D soap film) at high Re, where h is indeed seen to decrease with Re for a spatial velocity sequence. The near-laminar regime, where one expects h to increase with Re, is not experimentally accessible.

4.6.1 Analysis

It is worthwhile at this point to give an overview of how the experimental data will be treated. For details on how the data were measured in the first place, refer back to Sec. 4.3. Several ideas already introduced in earlier sections are repeated here.

4.6.1.1 Short review of data treatment In the discussion that follows an uppercase U denotes the data (the random variable, the message) with possible velocity values $\mathcal{U}$ and the lowercase u denotes a particular member of that set. We can also consider groups of length L denoted by the set $\mathcal{U}^{(\mathcal{L})}$ and its particular members $u^{(L)}$. We are interested in treating a group because of the correlations that may exist

between its members. Overhead arrows indicate a direction in the 1D data set relative to an arbitrary reference point x. For example, $\overrightarrow{U^{(L)}}$ refers to any block of data of size L taken to the right of x. For example, if $L = 3$, then a particular block $\overrightarrow{u^3}$ is as below

$$...u_{x-\Delta x}, u_x, \overrightarrow{u_{x+\Delta x}, u_{x+2\Delta x}, u_{x+3\Delta x}}, u_{x+4\Delta x}, ...$$

where Δx is the spatial resolution. If no L is mentioned, the block is (semi-)infinite.

Let U be a velocity component in the soap film, which is characterized by the experimental probability distribution $P(U)$. The focus is on the information shared between different directions $\overleftarrow{U}$ and $\overrightarrow{U}$ relative to the arbitrary point x [19, 5]. If we had data with explicit time dependence, we would talk about the past, future and present [5].

4.6.1.2 Partitions In order to use our information formalism, the experimental data must be converted to symbols [28]. A partition is defined which separates the data into disjoint slices of size ϵ. The data values in each specified range (slice) are then assigned to a unique symbol [28, 26]. That is, if data points are ϵ apart, they correspond to different symbols. (In some sense all experiments do this because of their limited precision.) The size and location of the divisions can be chosen to faithfully represent the original system even for seemingly coarse partitions [28].

Figure 37 shows an example of how a partition is used for some of the turbulence data where the mean speed has been subtracted out and the result divided by the rms velocity. The distance between the horizontal dividers is the partition size ϵ. Data falling in between the same walls (or on the boundary as prescribed) are assigned to the same symbol. The number of symbols required to describe the data is inversely related to the magnitude of ϵ. For all the data used here, if $\epsilon \gtrsim 10$ and a divider line kept at 0, then the data are effectively binarized.

Correctly identifying those partitions which completely describe the system (called "generating") can be extremely difficult, if they even exist [26]. However, much can and has been learned about complex systems such as the brain or turbulence even after converting a data series into a simple binary alphabet [28, 29, 105, 30]. Approximate treatments are usually necessary and often useful, as long as they still represent the underlying system [28]. For a chaotic time series, the entropy rate h may approach a constant value (h_{KS}) as ϵ decreases [106, 26]. For a spatially extended system, one may expect a similar asymptotic behavior.

Now let us look at how this partition might affect our data. We take some raw turbulence data as seen in Fig. 38. We can calculate the mean $\langle u \rangle$ and the rms $u' = \sqrt{\langle u^2 \rangle}$ with little difficulty. In order to treat data in a similar way (and so that they have roughly the same number of possible symbol values), we subtract out the mean and divide by the rms. Then, we have a non-dimensional data set and we divide the data into sets like the ones shown above in Fig. 37.

We can see from Fig. 38 that there is hardly any difference between the raw data and the partitioned data when the partition is this small ($\epsilon = 0.1u'$). This small of a partition is very nearly the smallest we

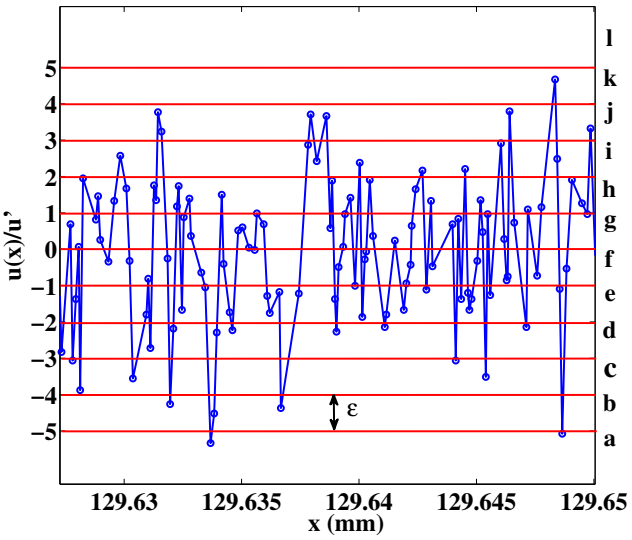


Figure 37: An example of a partition used with turbulence data. The dividers are a distance $\epsilon = 1$ apart. Data points that fall inside a section are assigned to that section's symbol. As ϵ decreases, the number of symbols required to describe the data increases.

can use. The reason for this is that smaller partition sizes would mean we need more data to accurately determine the probabilities of large blocks of data $U^{(L)}$. The next partition size $\epsilon = 0.5u'$ poses also follows the data quite faithfully. A large partition $\epsilon = 1u'$ also follows the general trends, but is clearly distinct from the raw data.

The general effect of partitioning is to filter out small velocity motions (not spatial filtering!). One could view this as looking at the large eddies in the turbulent flow and ignoring the small ones. This is similar to spatial filtering in the sense that all eddies of a particular size also have a characteristic velocity magnitude. With this in mind, even binarizing the data as in Fig. 41 is not stupid since we will retain some influence from the largest scale motions. We show in Sec. 4.6.2 that the general behavior of h is independent of the partition.

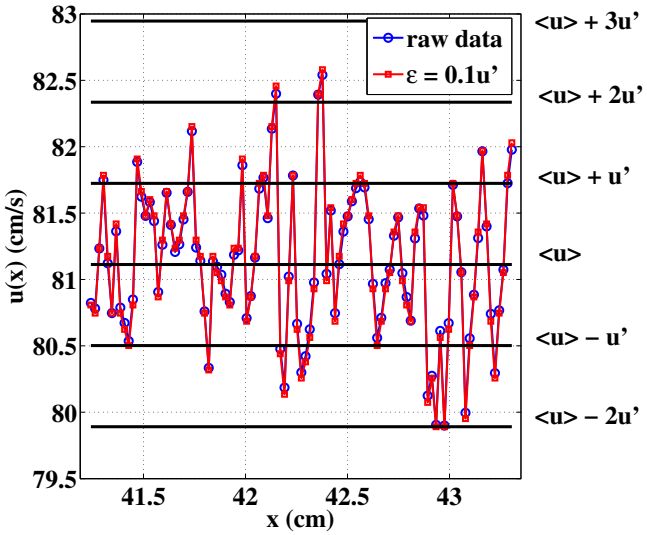


Figure 38: Raw turbulence data for low $Re \simeq 20$. The rms velocity is $u' \simeq 0.85$. A partition is used which distinguishes data that are $\epsilon = 0.1u'$ apart. The data are then replotted on top of the original raw data to show the difference. A partition at $u = 0$ is the reference partition.

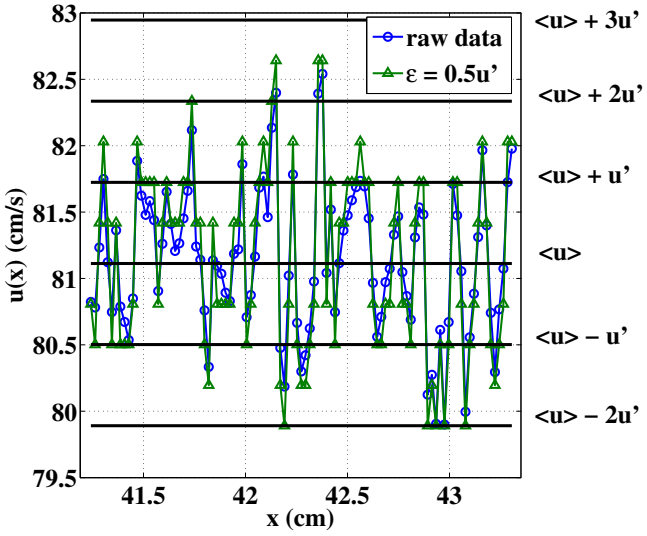


Figure 39: Raw turbulence data for low $Re \simeq 20$. The rms velocity is $u' \simeq 0.85$. A partition is used which distinguishes data that are $\epsilon = 0.5u'$ apart. The data are then replotted on top of the original raw data to show the difference. A partition at $u = 0$ is the reference partition.

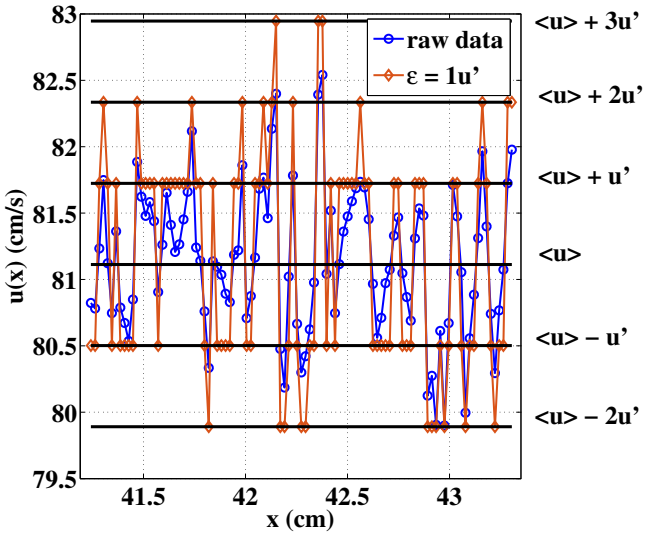


Figure 40: Raw turbulence data for low $Re \simeq 20$. The rms velocity is $u' \simeq 0.85$. A partition is used which distinguishes data that are $\epsilon = 1u'$ apart. The data are then replotted on top of the original raw data to show the difference. A partition at $u = 0$ is the reference partition.

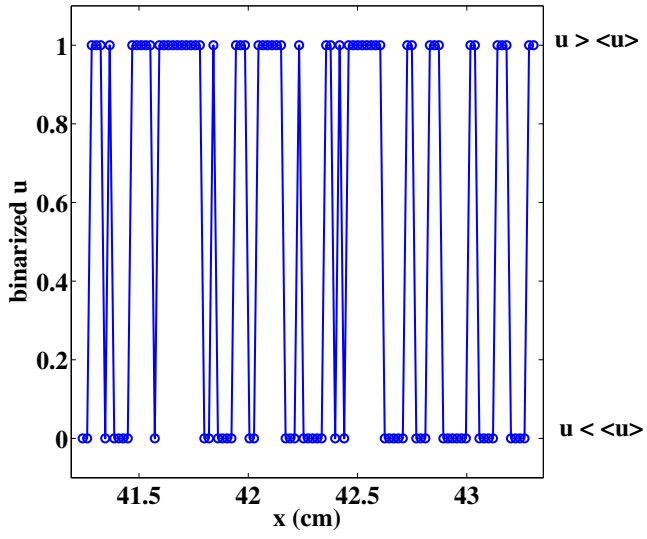


Figure 41: Raw turbulence data for low $Re \simeq 20$. A partition is used which distinguishes data that are above and below the mean value of the velocity. This effective binarizes the data. Refer back to Figs. 38-40 to see the original raw data. The general trend of the "large scales" is still present.

A simple check to see if one's partition is an appropriate size is to look at the PDFs of the partitioned data as done in Fig. 42. If the partition is too small, then the PDF will look blocky and not smooth. If the partition is too large, then there is not enough data to resolve the probabilities for the partitions and the PDF may begin to oscillate [123]. A good rule of thumb is that for turbulent velocity data, one should have at least 3 and better close to 10 partitions per rms division [123]. This was the criterion used for an interesting stochastic analysis of turbulent jet data [124, 125, 123].

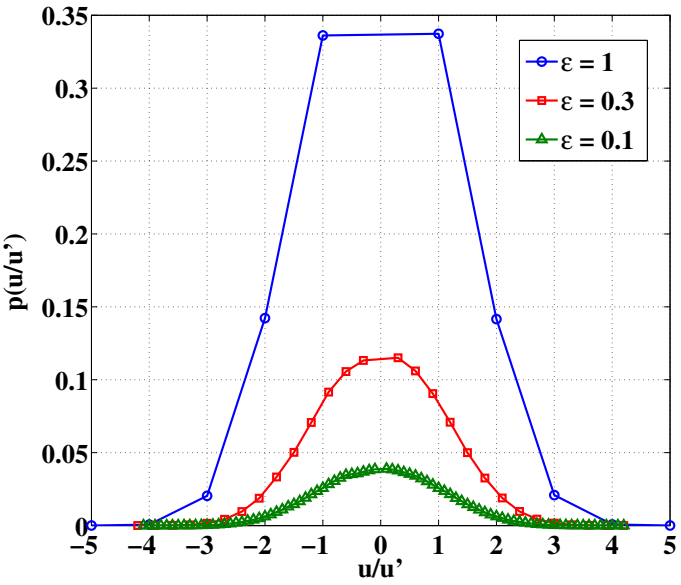


Figure 42: PDFs of the velocity for different partition sizes $\epsilon = 1$ (◯), 0.3 (□), and 0.1 (△). The enstrophy data for $Re \simeq 6000$ were first normalized by their rms value u'. The overall values of the probability change for different ϵ, as is expected [4], but the form of the PDFs is essentially the same.

4.6.1.3 Blocks of data Once the data have been partitioned, we can begin our processing of the data. It is important to recognize that the partitioning of the data involves the resolution of u, while blocking the data as we will soon do involves the spatial resolution (x). Consider the partitioned data from Fig. 40 from above. The first 10 data points are

$$... - 1, -1, +1, +2, +1, -1, +1, -1, -1, -1, ...$$

where -1 means this velocity value was originally between $u = 0$ and $u = -1u'$. Now suppose we are interested in calculating the probabilities of blocks of data that are $L = 3$ in size. We will count up all the blocks that we record and the probabilities will be estimated as this number divided by the total number of $L = 3$ blocks that appear. In this substring of 10 data points, the first 4 blocks we see are

$$... - 1, -1, +1, ...$$

$$\ldots -1, +1, +2, \ldots$$

$$\ldots +1, +2, +1, \ldots$$

$$\ldots +2, +1, -1, \ldots$$

Notice that we are sliding along by one data point at a time. The second $L = 3$ block we considered has two data points from the first block and so on. Because our measured data set is assumed to be stationary, we have no reference point and can look at any block at any position. Indeed, we must[13]. Keep in mind, however, that when we calculate conditional probabilities between blocks of $L > 1$, the two blocks are adjacent but do not share any data points.

4.6.2 Results on h for 2D turbulence

The main results on h appear in Fig. 43, which is a plot of h (see Eq. 2.3) and c (see Eq. 2.4) vs. Re, where $Re \equiv u'w/\nu$ and ν is the kinematic viscosity. The open circles ($\bigcirc$) and squares ($\square$) show h and c respectively for the binary partition. The two upper data sets ($\triangle$,$\triangleright$) are h and c for a finer partition where the velocity data are distinguished by their first significant figure ($\epsilon = 1$). The same trend is shown by both partitions and for all partitions studied, namely that h (and c) are decreasing functions of Re.

Note that h and c are only weakly dependent on Re: $h \propto -\log Re$ + const. after an initial plateau. The decrease begins as soon as a cascade appears, as seen in Fig. 43. The decrease is independent of the type of cascade, as both the energy and enstrophy cascade data are present in the figure. The flat region at low Re corresponds to the data without a cascade. Representative spectra for these regions can be seen in Fig. 24.

At first glance this result seems surprising. We are used to thinking of turbulence as the prototypical unpredictable system. (Some even use atmospheric turbulence as a random number generator [131].) This is, of course, true if we look at temporal variations [129, 130]. Here we see a fundamental distinction between the temporal and spatial features of turbulence.

However, the decrease is in accord with the common picture of the turbulent cascade [35, 32, 33]. The energy (enstrophy) flows from one scale r to nearby spatial scales. The eddies participating in the cascade are necessarily correlated and the extent of the inertial range (cascade region) increases with Re.

Although the system under study here is two dimensional, the same decrease in h should also hold for three dimensional turbulence in the fully developed regime. It should be noted that the results in Fig. 43 appear to be somewhat similar to that of Wijesekera *et. al.* in their study of spatial density fluctuations in the ocean [118].

We can understand the decrease in h by looking at correlations, which are becoming increasingly important as Re increases. This is evidenced by the decrease in the (spatial) decay rate γ_C for the mutual information $I(\Delta)$ as shown in Fig. 44. Note the similarities between Fig. 44 and Fig. 43. The increased strength of the correlations is responsible for the decrease in h.

[13] A practical reason for wanting to do this is so that we can get better estimates of the probabilities.

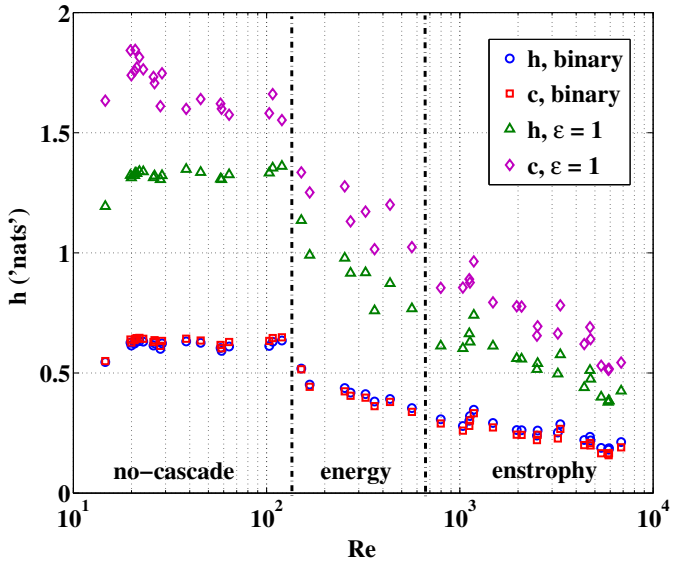


Figure 43: Entropy density h and compression ratio c vs. Re for the 2D turbulent data. h is a decreasing function of Re. The flat, higher h region corresponds to the no-cascade data. The decay begins with the emergence of the cascade. A binary partition (h: ◯, c: □) and a second partition where the data is saved to the nearest integer (h: △, c: ⋄) are shown here. Dividing lines are shown that separate the data into no-cascade, energy and enstrophy regions respectively.

We will continue painting our information picture of turbulence using C and E in the next section. Here we simply emphasize that (1) the spatial and temporal structure of turbulence evolves differently with Re, (2) turbulence is less spatially unpredictable as Re increases and (3) this Re dependence comes only after a cascade develops (see Fig. 43).

4.6.3 A note on partitions

The value of h is sensitive to the size of the partition ϵ when converting to symbols as shown in Fig. 45. Here h is plotted as a function of ϵ for three different values of Re. Although h increases as ϵ decreases, the curves never cross for the different Re. The reason for the general inverse relationship between ϵ and h is that as ϵ decreases, more detailed information is described by the symbols. The location of the dividers is important for the coarser partitions [115], but as the partition size decreases the results are not sensitive to this placement.

The compression ratio c is not a reliable estimate of h at the finer partitions, but it is still an indicator of the information content of the data streams and also shows the same decrease with Re [111]. It is also not at all surprising that h has a significant ϵ dependence. The important point here is that at any ϵ the behavior is the same.

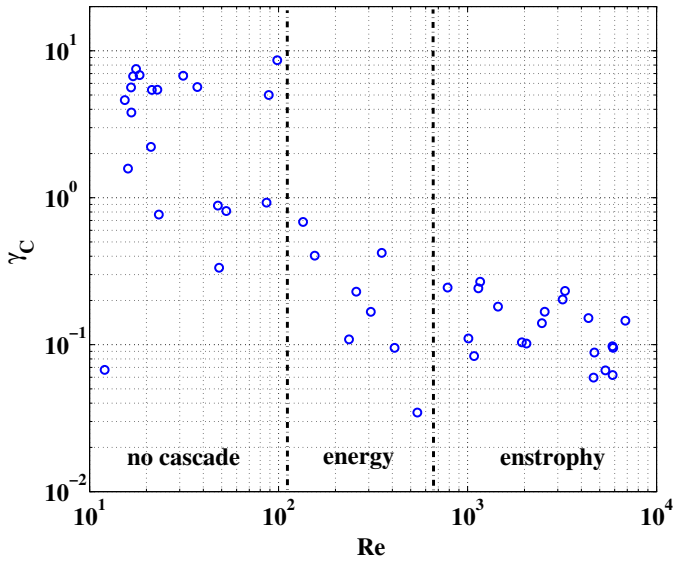


Figure 44: Spatial decay rate of mutual information for turbulence data using a partition where data are saved to the nearest integer. The decay rate γ_C is calculated as the reciprocal of the area underneath the mutual information curve: $1/\sum_0^\infty I(\Delta)$. If this is large, then the correlations are weak. Dividing lines are shown that roughly separate the data into no-cascade, energy and enstrophy regions respectively. As Re increases, the correlations get stronger.

An estimate can be made for the smallest size of the partition needed to resolve all fluctuations in the entire inertial range, based on the smallest eddy's characteristic velocity u_η [35, 119, 120]. Simple estimates show that the smaller partitions are fine enough to resolve a fluctuation of this magnitude for all Re. It is surprising that even a binary partition captures the main features: $h \propto -\log Re + \text{const}$. This suggests that one may fruitfully study turbulence just by looking at these 1s and 0s. Similar studies of complex systems such as the brain and heart and even turbulence [28, 105, 30, 29] have also used very coarse partitions.

We cannot avoid partitioning, but as an additional verification of the validity of this coarse-graining approach, the decay rates calculated from the raw data using the autocorrelation method and with the mutual information were compared for various partitions. The Re-dependence was almost exactly the same, although there is a shift by a factor of $1/e$ for the mutual information method. Since the entropy is fundamentally connected to correlations, this is strong evidence for the validity of this coarse-graining approach.

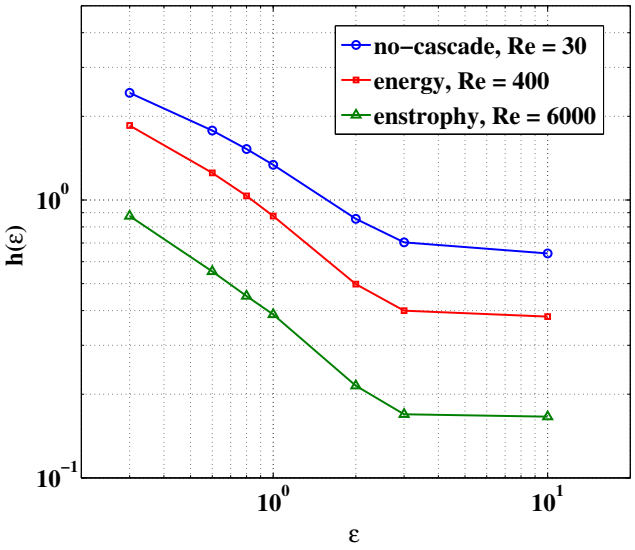


Figure 45: Entropy density h as a function of the partition size ϵ for three different Re. (The largest partition size corresponds to binarized data.) The three curves correspond to the no-cascade (◯), energy cascade (□) and enstrophy cascade (△) data. Despite the significant ϵ-dependence, none of the curves intersect. This means that the Re-dependence is ϵ-independent.

4.7 COMPLEXITY AND PREDICTABILITY

Now we include E and C in our analysis of the behavior of predictability in 2D turbulence. The quantities C, E and h are plotted vs. Re in Fig. 46 for $A = 2$ (binarized data). The data are roughly divided in Re into no-cascade (flat $\mathcal{E}(k)$ for $Re < 80$) and cascade (power law $\mathcal{E}(k)$ for $Re > 80$) regimes. Although C and E intersect at finite $Re \simeq 7000$ in Fig. 46, for larger alphabets of size $A > 2$ we find a different meeting point. However, the Re-dependent behavior of h, E and C discussed below is the same for all partitions tested (up to $A = 8$).

4.7.1 Cascade Turbulence

Now consider the behavior of h, E and C (see Eqs. 2.3, 2.12 and 3.1) in the "cascade regime" of Fig. 46, $Re > 80$. At these values of Re, $\mathcal{E}(k)$ shows power law scaling as in Fig. 24. Both energy and enstrophy cascade data are present. We see from Fig. 46 that the unpredictability (h) is decreasing, the amount of information needed to predict (C) is also decreasing, while the reduction in unpredictability (E) is increasing (all logarithmically). The opposite trend in Re for E and C is noteworthy. It is surprising that the behavior of h, E and C for $Re > 80$ does not depend on which cascade is present, only on whether or not there is a

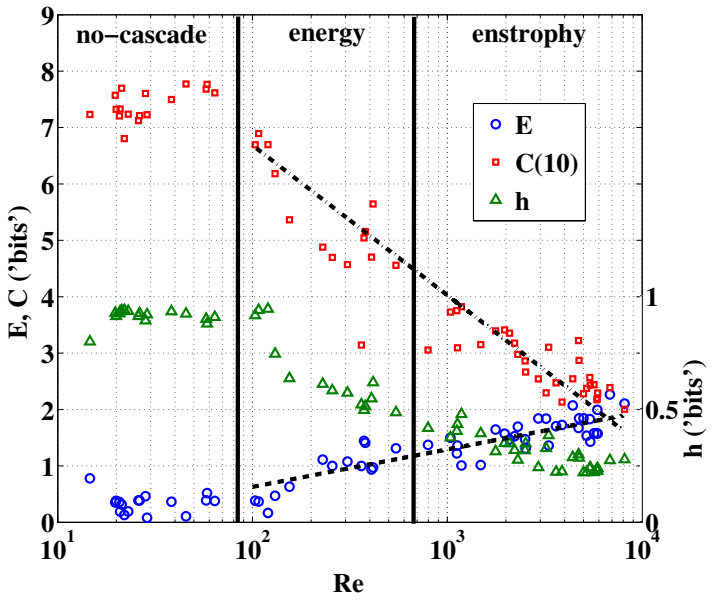


Figure 46: The statistical complexity C (□), excess entropy E (◯) and entropy density h (△) as a function of Re for binarized ($A = 2$) data. We plot h on a different scale for better visibility. The maximum value of h here is $\log_2 2 = 1$, which the no-cascade data for $Re < 80$ approach very closely. Here $L = 10$ and we used our MATLAB program with the χ^2 test to calculate C. The lines are not fits to the data but are meant to suggest the behavior of C and E as functions of Re. For the cascade region, C and h are decreasing functions of Re while E increases. The vertical lines roughly separate the data according to their cascade.

cascade at all.

The increase of E with Re can be understood from the traditional view that as Re increases, the inertial range of correlated scales grows. As a result, more information is shared between distant points across longer scales. Dimensional arguments suggest that the turbulent degrees of freedom go as $N \propto Re$ for the enstrophy cascade and $N \propto Re^{3/2}$ for the inverse energy cascade. In the 3D energy cascade, $N \propto Re^{9/4}$ [134]. Thus the behavior $E \propto \log_2 Re$ in Fig. 46 indicates that E is a logarithmic measure of the extent of the inertial range.

An interpretation of the behavior of C is also suggested by the traditional picture of 2D turbulence. As Re grows, the inertial range broadens, and more of the velocity fluctuations come under the governance of the cascade. Thus, the randomness h will decrease, and because the cascade's structure is dominating, our prediction cost C decreases. This is the result of the general principle that patterns help us to predict [12]. Here the pattern is the cascade's structure.

With h, E and C we see that the spatial predictability of (2D) turbulence is increasing with Re in its fullest sense: we can predict further and more easily. This is in stark contrast to turbulence's increasing

temporal unpredictability with Re [129, 130]. This reiterates the important difference between time and space in turbulence, which is of fundamental interest and practical importance (recall the airplane).

Turbulence has traditionally been thought of as unpredictable [40, 43]. Usually, this is to contrast it with laminar flow. With h, E and C, we can not only quantify this but see how the degree of predictability depends on Re. While we have argued for a qualitative understanding of these quantities, a quantitative analysis is still necessary. The largest mystery is the difference $\chi = C - E$. In principle, one should be able to understand why the correlations accounted for by E are not enough to make an accurate prediction.

4.7.2 Transition to Cascade Turbulence

Next consider the region of Fig. 46 labeled "no-cascade". The absence of a cascade is evidenced by no power law scaling in $\mathcal{E}(k)$ as in Fig. 24. Here h, E and C are relatively constant with respect to Re. It is notable that h is very near to the random (white noise) value of $\log_2 2 = 1$, which is nothing like laminar flow where $h = 0$. When a cascade emerges at $Re \simeq 80$, all three quantities begin to change noticeably. This change in behavior is decidedly different from the laminar to turbulent transition which only involves the onset of fluctuations [40].

The fluctuations of pre-cascade turbulence are apparently difficult to predict (C is large in Fig. 46). Moreover, the wide separation between E and C is surprising. We emphasize that C, E and h have made a clear distinction between simply unsteady velocity fluctuations and cascade turbulence. It is natural that tools designed to quantify randomness and order should be able to detect this transition.

Simulations of turbulence have shown that statistics of the velocity derivatives are gaussian (or sub-gaussian) up until a small value of the Taylor microscale Reynolds number $Re_\lambda \simeq 10$ [132]. Below this value of Re_λ, there is, in their words, a "regime which is a complex time-dependent flow rather than a turbulent one." The transition to what we have called here "cascade turbulence" is there evidenced by non-gaussian velocity derivative statistics and scaling in the structure functions. Recall that nongaussian statistics are a general feature of fully developed turbulence [146].

We also resort here to a more traditional tool from turbulence: the correlation function $c(r) \equiv \langle u(x)u(x+r)\rangle_x / u'^2$. This is plotted in Fig. 47. $c(r)$ has typically been thought of as a tool for determining the range of length scales over which u is correlated. We see that in Fig. 47 $c(r)$ is telling us that for small Re, there is nearly no range of scales over which u is correlated. Of course, $c(r)$ does not address our principle question of predictability (nor is it a perfect measure of correlation [142, 141]).

Figures 24 and 47 both seem to indicate that for $Re \leq 100$ the data is like white noise, as does h and E from Fig. 46. However, if this were the case then C should also be zero in this regime, which it is not. There is something peculiar happening in this regime and something very interesting occurs such that this no-cascade, non-laminar flow should suddenly develop structure and order.

It seems reasonable to draw an analogy with the insight of Kadanoff on Rayleigh-Benard systems [148]. In his assessment of the general behavior of these systems, where a layer of fluid is heated from below,

he suggested that the reason structure is observed (and ultimately a turbulent regime) is that the system organizes itself to take heated fluid particles from the bottom of the layer to the top as quicly as possible. Perhaps a similar thing is happening in this and other turbulent systems, where kinetic energy is injected at some large scale and the system organizes itself to dissipate this energy as quickly as possible, thus forming a cascade.

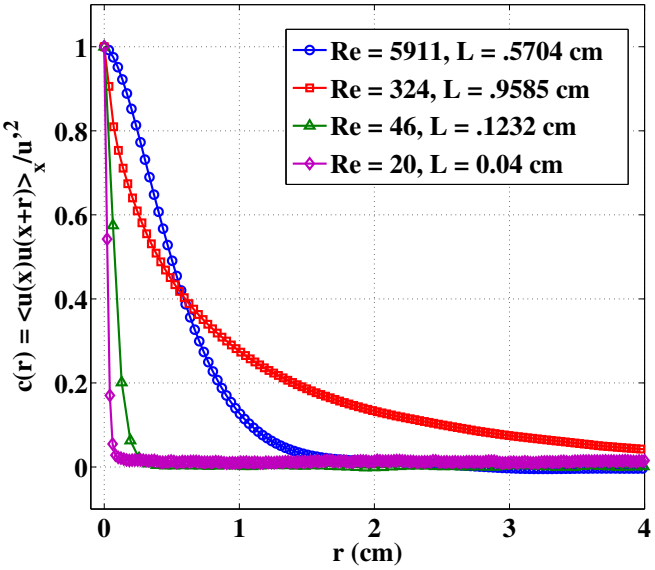


Figure 47: The velocity autocorrelation function $c(r)$ plotted *vs.* r for several values of Re. For small Re, $c(r)$ quickly decays to zero, indicating little correlation in the velocity u. For larger Re, where Fig. 24 indicates spatial structure, there is a wider range of correlated scales. The $Re = 324$ curve has a longer correlation length L than the higher $Re = 5911$ curve presumably because this lower Re curve corresponds to an inverse energy cascade. The inverse energy cascade is supposed to involve larger length scales than the enstrophy cascade.

The traditional approaches to the laminar-turbulent transition deal with instabilities of the laminar flow. Whether it is the quasi-periodicity of Landau [134] or the nonperiodicity of Ruelle and Takens [144], none of these approaches deal with the development of a cascade [145]. And yet a cascade is always present in fully-developed turbulence [68, 35]. How does this arise? New approaches and models are necessary to understand how cascade behavior develops out of the "complex, time-dependent flow" [132]. Since this development is clearly visible in Fig. 46, an information theory approach seems promising.

We suggest an information-theoretic indicator of a cascade. Based on the above arguments, large E and $1/C$ should both indicate a well-developed cascade. With that in mind, we can also consider the "predictive efficiency" E/C [37], which is an increasing function of Re, as shown in Fig. 48 for two different partitions. The ratio E/C tells us the fraction of the information needed to predict C that is due to correlations E. It is nearly zero when no cascade is present and grows smoothly after one has emerged. This shows that E/C

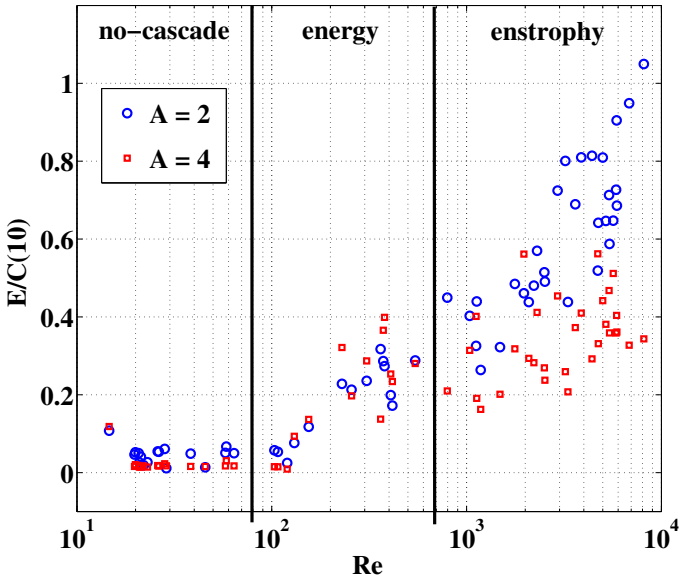


Figure 48: The predictive efficiency E/C plotted vs. Re using the same data as in Fig. 46 as well as a quaternary partition $A = 4$ with partitions placed symmetrically with respect to the mean. We used $L = 10$ for both partitions. Here we find that E/C is increasing only after a cascade develops.

is a nice tool for studying the transition to cascade turbulence.

Besides this cascade transition, the laminar to fluctuation transition is also of interest. Unfortunately, we are not able to access a truly laminar regime with our apparatus. For laminar flow and this geometry, $h = E = C = 0$ [5]. Looking at Fig. 46, and with the reasonable assumption that h and C are continuous functions of Re, one expects a local maximum in C and h at some low value of Re. This maximum would correspond to a special transition in the evolution of the flow between laminar and turbulent behavior. The observation of this maximum requires a different experimental setup.

5.0 CONCLUSION

We have used the tools of information theory and computational mechanics to study turbulence. The primary function of these tools is to test the unpredictability and predictability of a system. We did this with h, E and C for many experimental velocity data sets taken from a 2D turbulent soap film. We have found a general Re-dependent behavior (after a cascade develops) which suggests that the spatial features of turbulence are behaving in a way not previously considered. The unpredictability is actually decreasing with increasing Re, and at the same time our ability to make predictions is increasing. This is a counter-intuitive result. However, it has added to our understanding of turbulence by clearing up some common confusion about its behavior at large Re.

The most surprising finding is the existence of a second transition in turbulence. The usual laminar to turbulence transition is not explored here. Instead, we observe a special transition in which structureless fluctuations become cascade turbulence. This transition can be detected with traditional measures, but it is most clear with the information tools used here.

We have also introduced some new measures and tools for studying intermittency. In 2D turbulence we have seen that the intermittency (by our measures) has a strong Re-dependence. This aspect has not been noticed or discussed before now. We have tried to make a connection with the intermittency observed and the basic assumptions that lead us to call it intermittency in the first place. Namely, do the small scales really forget the large scales and by so doing become universal? We have made an attempt at testing this using very basic tools from information theory.

In summary, our information bag of tools are appropriate for answering questions people often ask about turbulence. Is it predictable? Are the small scales universal? Rather than leaving these questions answered only qualitatively, we have attempted quantitative answers.

BIBLIOGRAPHY

[1] N. Gershenfeld, *The Physics of Information Technology* (Cambridge University Press, Cambridge, 2000)

[2] C. E. Shannon, "A mathematical theory of communication," Bell. Sys. Tech. Journal **27**, 379-423, 623-656 (1948)

[3] C. E. Shannon, W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1964)

[4] T. M. Cover, J. A. Thomas, *Elements of Information Theory, 2nd Ed.* (Wiley, New York, 1991)

[5] J. P. Crutchfield, Nat. Phys. **8**, 17 (2012)

[6] L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1962)

[7] E. T. Jaynes, Phys. Rev. **106**, 620 (1957)

[8] E. T. Jaynes, Phys. Rev. **108**, 171-190 (1957)

[9] E. T. Jaynes, Proc. IEEE **70**, 939 (1982)

[10] R. K. Pathria, P. D. Beale *Statistical Mechanics, 2nd Edition* (Elsevier, Boston, 2011)

[11] J. P. Crutchfield, K. Young, Phys. Rev. Lett. **63**, 105 (1989)

[12] C. R. Shalizi, J. P. Crutchfield, J. Stat. Phys. **104**, 817 (2001)

[13] S. Lloyd, Nature **406**, 1047 (2000)

[14] R. Landauer, Phys. Lett. A **217**, 188 (1996)

[15] R. Landauer, Physics Today, 25 (1991)

[16] R. Quax, A. Appolloni, P. M. A. Sloot, Eur. Phys. J. Spec. Top. **222**, 1389 (2013)

[17] R. S. Shaw, Z. Naturforsch **36A**, 80 (1981)

[18] R. T. Cerbus, W. I. Goldburg, Phys. Rev. E **88**, 053012 (2013)

[19] J. P. Crutchfield, D. P. Feldman, Phys. Rev. E **55**, R1239 (1997)

[20] J. R. Mahoney, C. J. Ellison, R. G. James, J. P. Crutchfield, Chaos **21**, 037112 (2011)

[21] C. J. Ellison, J. R. Mahoney, J. P. Crutchfield, J. Stat. Phys. **136**, 1005 (2009)

[22] J. P. Crutchfield, C. J. Ellison, J. R. Mahoney, Phys. Rev. Lett. **103**, 094101 (2009)

[23] C. R. Shalizi, K. L. Shalizi, J. P. Crutchfield, arXiv:cs/0210025v3 [cs.LG]

[24] *http : //vserver*1*.cscs.lsa.umich.edu/ crshalizi/CSSR/*

[25] E. L. Lehmann, J. P. Romano, *Testing Statistical Hypotheses, 3rd Ed.* (Springer, New York, 2005)

[26] T. Schürmann, P. Grassberger, Chaos **6**, 414 (1996)

[27] J. P. Crutchfield, D. P. Feldman, Chaos **13**, 25 (2003)

[28] C. S. Daw, C. E. A. Finney, E. R. Tracy, Rev. Sci. Instrum. **74**, 915 (2002)

[29] A. J. Palmer, C. W. Fairall, W. A. Brewer, IEEE Trans. Geo. Remote Sensing **38**, 2056 (2000)

[30] M. Lehrman, A. B. Rechester, Phys. Rev. Lett. **87**, 164501 (2001)

[31] T. Tran, P. Chakraborty, N. Guttenberg, A. Prescott, H. Kellay, W. I. Goldburg, N. Goldenfeld, G. Gioia, Nat. Phys. **6**, 438 (2010)

[32] H. Kellay, W. I. Goldburg, Rep. Prog. Phys. **65**, 845-894 (2002)

[33] G. Boffetta, R. Ecke, Ann. Rev. Fluid Mech. **44**, 427 (2012)

[34] H. Kellay *et al*., Phys. Rev. Lett. **109**, 254502 (2012)

[35] P. A. Davidson, *Turbulence: An Introduction for Scientists and Engineers* (Oxford University Press, Oxford, 2004)

[36] D. P. Feldman, J. P. Crutchfield, Phys. Lett. A **238**, 244 (1998)

[37] K. Wiesner, M. Gu, E. Rieper, V. Vedral, Proc. Roy. Soc. A **468**, 4058 (2012)

[38] M. Tribus, E. C. McIrvine, Sci. Am., 224 (1971)

[39] http://www.gutenberg.org

[40] D. J. Tritton, *Physical Fluid Dynamics* (Oxford University Press, USA, 1988)

[41] H. P. Yockey, *Information Theory, Evolution, and the Origin of Life* (Cambridge University Press, New York, 2010)

[42] J. P. Eckmann, D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985)

[43] H. Tennekes, J. L. Lumley *A First Course in Turbulence* (The MIT Press, Cambridge, Mass.)

[44] W. Feller, *An Introduction to Probability Theory and Its Applications* (John Wiley and Songs, Inc., New York, 1950)

[45] H. Jeffreys, *Theory of Probability* (Oxford University Press, London 1939)

[46] L. Smith, *Chaos: a very short introduction* (Oxford University Press, New York, 2007)

[47] S. H. Strogatz, *Nonlinear Dynamics and Chaos (with applications to physics, biology, chemistry, and engineering)* (Perseus Publishing, Cambridge, MA, 1994)

[48] E. N. Lorenz, J. Atmos. Sci **20**, 130 (1963)

[49] L. F. Richardson, *Weather Prediction by Numerical Process* (Cambridge University Press, Cambridge, 1922)

[50] L. F. Richardson, Proc. Roy. Soc. A **110**, 709-737 (1926)

[51] U. Frisch, *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge University Press, UK 1995)

[52] G. I. Barenblatt, *Scaling* (Cambridge University Press, Cambridge, 2003)

[53] L. Onsager, Il Nuovo Cimento , 279-287 (1949)

[54] R. H. Kraichnan, Phys. Fluids **10**, 1417 (1967)

[55] G. K. Batchelor, Phys. Fl. **12**, 233 (2004)

[56] G. K. Batchelor, *The Theory of Homogeneous Turbulence* (Cambridge University Press, Cambridge, 1953)

[57] P. Chakraborty, T. Tran, G. Gioia, J. Elasticity **104**, 105 (2011)

[58] R. T. Cerbus, W. I. Goldburg, Phys. Fl. **25**, 105111 (2013)

[59] R. H. Kraichnan and D. Montgomery, Rep. Prog. Phys. **43**, 547 (1980)

[60] P. Tabeling, Phys. Rep. **362**, 1 (2002)

[61] S. Chen, R. E. Ecke, G. L. Eyink, M. Rivera, M. Wan, and Z. Xiao, Phys. Rev. Lett. **96**, 084502 (2006)

[62] S. Chen, R. E. Ecke, G. L. Eyink, X. Wang, and Z. Xiao, Phys. Rev. Lett. **91**, 214501 (2003)

[63] G. Boffetta, J. Fluid Mech. **589**, 253 (2007)

[64] G. Boffetta and S. Musacchio, Phys. Rev. E **82**, 016307 (2010)

[65] M. A. Rutgers, Phys. Rev. Lett. **81**, 2244 (1998)

[66] W. B. Daniel and M. A. Rutgers, arXiv:nlin/0005008

[67] C. H. Bruneau and H. Kellay, Phys. Rev E **71**, 046305 (2005)

[68] A. N. Kolmogorov, Dokl. Akad. Nauk. SSSR **30**, 299 (1941) (Proc. R. Soc. Lond. A **434** (reprinted))

[69] R. H. Kraichnan, Phys. Fluids **10**, 2080 (1967)

[70] K. R. Sreenivasan and R. A. Antonia, Ann Rev. Fluid Mech. **29**, 435 (1997)

[71] G. Boffetta, A. Celani, and M. Vergassola, Phys. Rev. E **61**, R29 (2000)

[72] J. Paret and P. Tabeling, Phys. Fluids **10**, 3126 (1998)

[73] Y. Jun and X. L. Wu, Phys. Rev. E **72**, 035302 (2005)

[74] G. Boffetta, A. Celani, S. Musacchio, and M. Vergassola, Phys. Rev. E **66**, 026304 (2002)

[75] K. Nam, E. Ott, T. M. Antonsen Jr., and P. N. Guzdar, Phys. Rev. Lett. **84**, 5134 (2000)

[76] Y. K. Tsang, E. Ott, T. M. Antonsen Jr., and P. N. Guzdar, Phys. Rev. E **71**, 066313 (2005)

[77] J. Paret, M. C. Jullien, and P. Tabeling, Phys. Rev. Lett. **83**, 3418 (1999)

[78] G. Boffetta, A. Cenedese, S. Espa, and S. Musacchio, Europhys. Lett. **71**, 590 (2005)

[79] A. Bracco and J. C. McWilliams, J. Fluid Mech. **646**, 517 (2010)

[80] R. Benzi, G. Paladin, S. Patarnello, P. Santangelo, and A. Vulpiani, J. Phys. A: Math. Gen. **19**, 3771 (1986)

[81] J. Jimenez, Proc. 15th 'Aha Huliko' Winter Workshop, Extreme Events, 81, (2007)

[82] H. Kahalerras, Y. Malécot, Y. Gagne, and B. Castaing, Phys. Fluids **10**, 910 (1998)

[83] W. Merzkirch, *Flow Visualization* (Academic, Orlando, 1987)

[84] A. Belmonte, B. Martin, and W. I. Goldburg, Phys. Fluids **12**, 835 (2000)

[85] A. Ramond and P. Millan, Expts. Fl. **28**, 58 (2000)

[86] S. A. Pope, *Turbulent Flows* (Cambridge University Press, Cambridge, 2000)

[87] A. Belmonte, W. I. Goldburg, H. Kellay, M. A. Rutgers, B. Martin, and X. L. Wu, Phys. Fluids **11**, 1196 (1999)

[88] P. Tabeling and H. Willaime, Phys. Rev. E **65**, 066301 (2002)

[89] C. V. Tran and D. G. Dritschel, J. Fluid Mech. **559**, 107 (2006)

[90] D. G. Dritschel, C. V. Tran, and R. K. Scott, J. Fluid Mech. **591**, 379 (2007)

[91] T. Dudok de Wit, Phys. Rev. E **70**, 055302 (1993)

[92] R. Benzi, R. Tripiccione, C. Baudet, F. Massaioli, and S. Succi, Phys. Rev. E **48**, R29 (1993)

[93] F. Anselmet, Y. Gagne, E. J. Hopfinger, and R. J. Antonia, J. Fluid Mech. **140**, 63 (1984)

[94] M. Rivera and X. L. Wu, Phys. Rev. Lett. **85**, 976 (2000)

[95] B. K. Martin, X. L. Wu, W. I. Goldburg and M. A. Rutgers, Phys. Rev. Lett. **80**, 3964 (1998)

[96] M. Mézard, A. Montanari, *Information, Physics, and Computation* (Oxford University Press, New York, 2009)

[97] Yu. L. Klimontovich, *Turbulent Motion and the Structure of Chaos* (Springer, Dordrecht, 1991)

[98] R. Kraichnan, Phys. Rev. Lett. **72**, 1016 (1994)

[99] B. I. Shraiman, E. D. Siggia, Nature **405**, 639 (2000)

[100] G. Falkovich, K. Gawedzki, M. Vergassola, Rev. Mod. Phys. **73**, 913 (2001)

[101] H. L. Swinney, J. P. Gollub, Physica D **18**, 448 (1986)

[102] A. Brandstäter, *et al.*, Phys. Rev. Lett. **51**, 1442 (1983)

[103] M. Yamada, K. Ohkitani, Phys. Rev. Lett. **60**, 938 (1988)

[104] D. Ruelle, Comm. Math. Phys. **87**, 287 (1982)

[105] M. Lehrman, A. B. Rechester, R. B. White, Phys. Rev. Lett. **78**, 54 (1997)

[106] E. Ott, *Chaos in Dynamical Systems* (Cambridge University Press, New York, 2002)

[107] D. P. Feldman, Information Theory Lecture Notes, http://hornacek.coa.edu/dave/Tutorial/notes.pdf

[108] D. P. Feldman, J. P. Crutchfield, Phys. Rev. E **67**, 051104 (2003)

[109] P. Grassberger, arXiv:physics/0207023v1

[110] D. Salomon, *Data Compression: The Complete Guide* (Springer-Verlag, London, 2007)

[111] A. Baronchelli, E. Caglioti, V. Loreto, Eur. J. Phys. **26**, S69 (2005)

[112] F. Kaspar, H. G. Schuster, Phys. Rev. A **36**, 842 (1987)

[113] W. Ebeling, T. Pöschel, A. Neiman, *Entropy and compressibility of symbol sequences,* in: PhysComp96, eds. T. Toffoli, M. Biafore and J. Leao (New England Complex Systems Institute, Cambridge, MA, 1996)

[114] E. I. Olivares, R. Vazquez-Medina, M. Cruz-Irisson, J.L. Del-Rio-Correa, *12th International Conference on Mathematical Methods in Electromagnetic Theory*, Odesa, Ukraine, 409 (2008)

[115] R. Steuer, L. Molgedey, W. Ebeling, M. A. Jiménez-Montaño, Euro, Phys. J. B **19**, 265 (2001)

[116] W. Ebeling, T. Poschel, K. F. Albrecht, Int. J. Bifurcation and Chaos **5**, 51 (1995)

[117] W. Li, J. Stat. Phys. **60**, 823 (1990)

[118] H. W. Wijesekera, T. M. Dillon, J. Geophys. Res. **102**, 3279 (1997)

[119] P. Gaspard, X. J. Wang, Phys. Rep. **235**, 291 (1993)

[120] X. J. Wang, P. Gaspard, Phys. Rev. A **46**, R300 (1992)

[121] K.R. Sreenivasan, G. Stolovitzky, J. Stat. Phys. **78**, 311 (1995)

[122] M. Rivera, W. Daniel, S. Chen, R. Ecke, Phys. Rev. Lett. **90**, 104502 (2003)

[123] J. Peinke, private communication.

[124] R. Friedrich, J. Peinke, Phys. Rev. Lett. **78**, 863 (1997)

[125] C. Renner, J. Peinke, R. Friedrich, J. Fluid. Mech. **433**, 383 (2001)

[126] G. H. Hardy, *A Mathematician's Apology*, (Cambridge University Press, New York, 1967)

[127] J. Earman, J. Norton, Stud. Hist. Phil. Mod. Phys. **29**, 435 (1998)

[128] J. Earman, J. Norton, Stud. Hist. Phil. Mod. Phys. **30**, 1 (1999)

[129] E. Aurell, et. al., Phys. Rev. E **53**, 2337 (1996)

[130] C. E. Leith, R. H. Kraichnan, J. Atmos. Sci. **29**, 1041 (1972)

[131] www.random.org

[132] J. Schumacher, K. R. Sreenivasan, V. Yakhot, New. J. Phys. **9**, 89 (2007)

[133] V. Yakhot, arXiv:1308.4205 [physics.flu-dyn]

[134] L. D. Landau, E. M. Lifshitz, *Fluid Mechanics* (Butterworth-Heinemann, 2nd Ed., Oxford, 1987)

[135] Boeing, http://www.boeing.com/boeing/commercial/... ...747family/pf/pf_400_prod.page, 4 July, 2014

[136] J. D. McMinn, AIAA Guidance Navigation and Control Conference, AIAA973532, (1997)

[137] F. Cecconi, M. Cencini, M. Falcioni, and A. Vulpiani, Am. J. Phys. **80**, 1001-1008 (2012).

[138] G. I. Barenblatt, *Scaling* (Cambridge University Press, Cambridge, 2003)

[139] D. Pisarenko, L. Biferale, D. Courvoisier, U. Frisch, and M. Vergassola, Phys. Fl. **5**, 2533-2538 (1993)

[140] M. Materassi, G. Consolini, N. Smith, and R. De Marco, Entropy **16**, 1272-1286 (2014)

[141] W. Li, J. Stat. Phys. **60**, 823 (1990)

[142] A. M. Fraser, H. L. Swinney, Phys. Rev. A **33**, 1134 (1986)

[143] G. L. Baker, J. P. Gollub, *Chaotic Dynamics: An Introduction* (Cambridge U. Press, 2nd Ed., Cambridge, 1996)

[144] D. Ruelle, F. Takens, Commun. Math. Phys. **20**, 167 (1971)

[145] H. L. Swinney, J. P. Gollub, Phys. Today **31**, 41 (1978)

[146] K. R. Sreenivasan, R. A. Antonia, Ann. Rev. Fl. Mech. **29**, 435 (1997)

[147] R. Landauer, Phys. Lett. A **217**, 188 (1996)

[148] L. P. Kadanoff, Phys. Today **54**, 34 (2001)

[149] M. Cencini, J. Bec, L. Biferale, G. Boffetta, A. Celani, A. S. Lanotte, S. Musacchio, F. Toschi, J. Turbulence **7**, 1 (2006)

[150] J. Lyklema, *Fundamentals of Interface and Colloid Science* (Academic Press, New York, 1993)

[151] Lecture Notes: $http://www.owlnet.rice.edu/\ phys534/notes/week10_lectures.pdf$

[152] E. A. Codling, M. J. Plank, S. Benhamou, J. R. Soc. Interface **5**, 813 (2008)

[153] R. Piazza, *Soft Matter* (Springer-Verlag, Dordrecht, 2011)

[154] $http://www.dantecdynamics.com/measurement-principles-of-lda$